

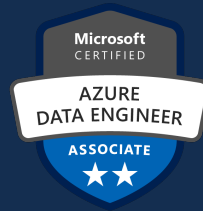
# Learn Azure Data Factory From Scratch



# About Me



Ramesh Retnasamy  
Data Engineer/ Architect



<https://www.linkedin.com/in/ramesh-retnasamy/>

# About this course



Azure Data Factory (ADF)

## Azure storage solutions



Azure SQL Database



Azure Blob Storage



Azure Data Lake Storage Gen2

## Other Bigdata Solutions



Azure Databricks

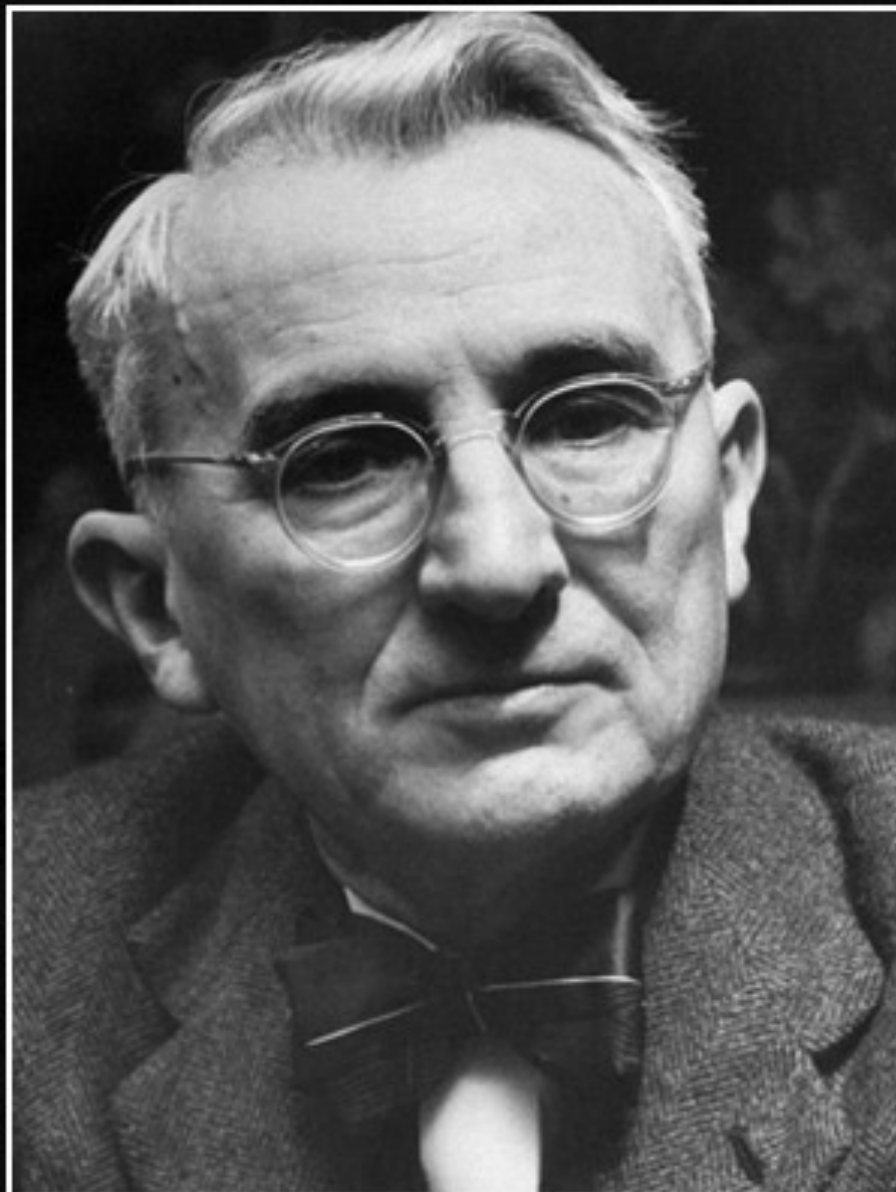


Azure HDInsight

## Reporting Technologies



PowerBI

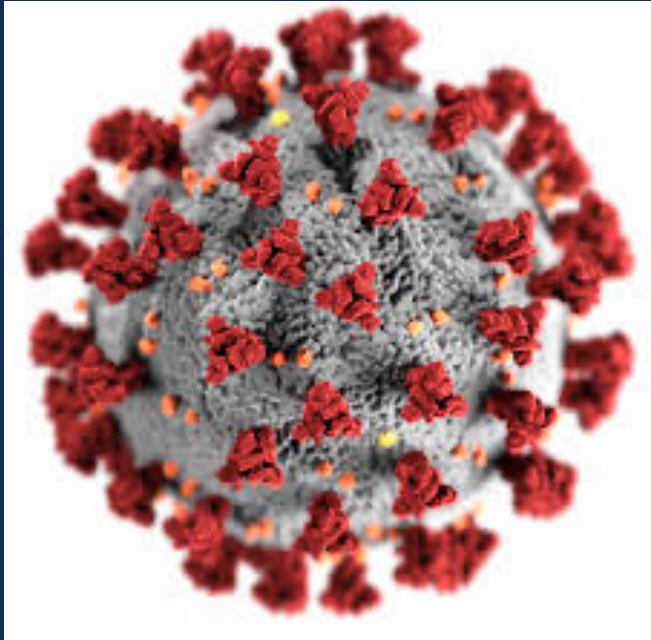


Learning is an active process. We learn by doing.. Only knowledge that is used sticks in your mind.

— Dale Carnegie —

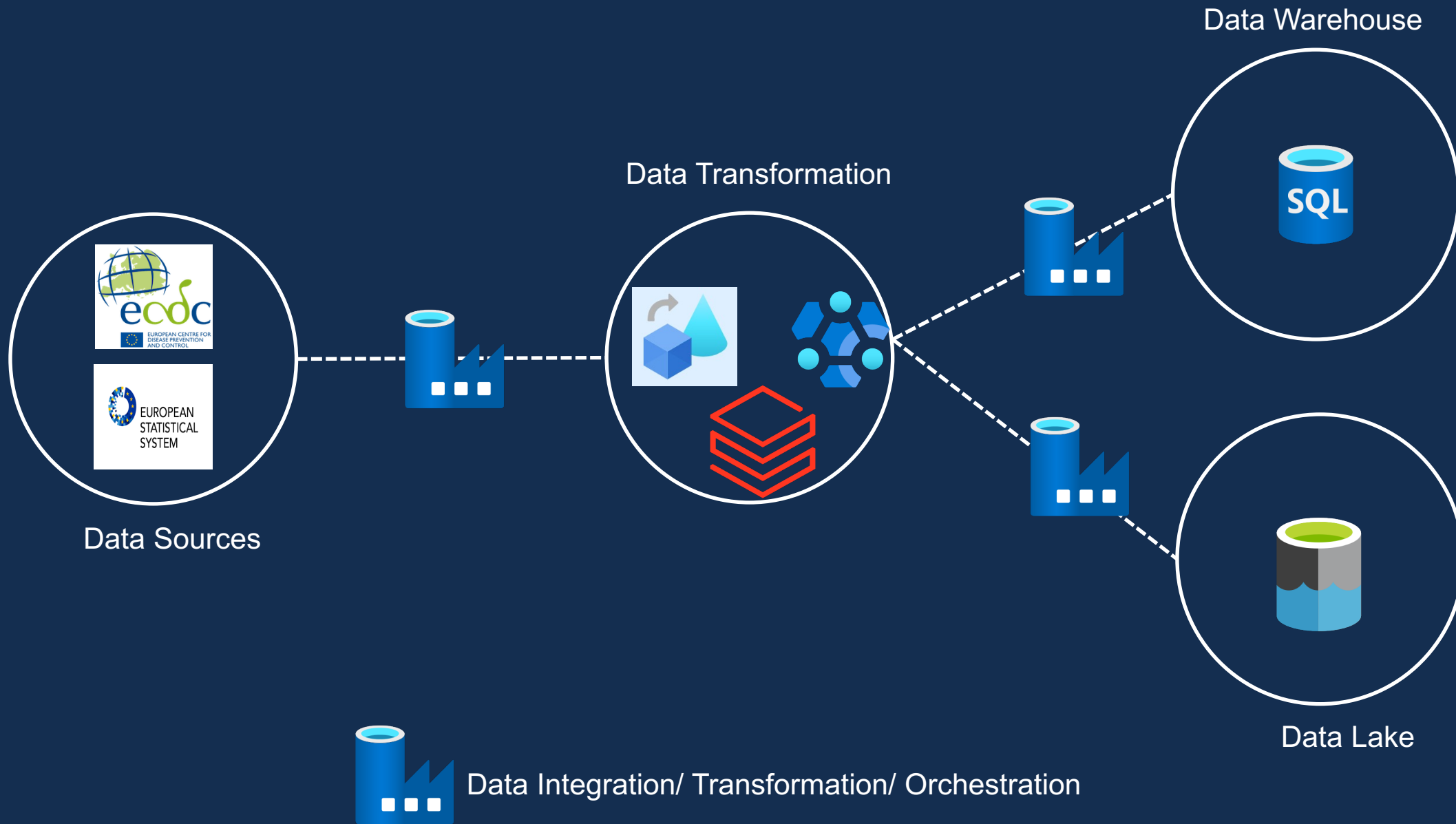
AZ QUOTES





# Covid-19 Prediction/ Reporting

# Covid-19 Prediction/ Reporting



# Covid-19 Prediction/ Reporting

## Covid-19 Cases EU/EEA & UK

Total Confirmed Cases

1,438,022

Total Deaths

31,981

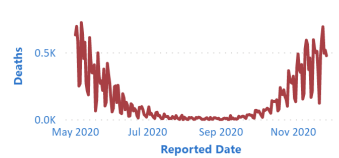
Reported Date (From - To)

01/05/2020 29/11/2020

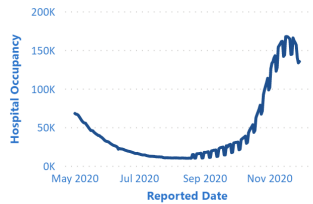
Total Cases Trend



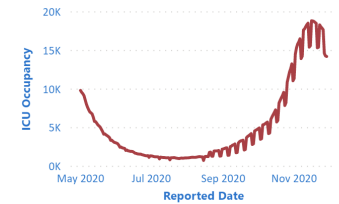
Total Deaths Trend



Hospital Occupancy Trend



ICU Occupancy Trend



Country

- ☐ Albania
- ☐ Andorra
- ☐ Armenia
- ☐ Austria
- ☐ Azerbaijan
- ☐ Belarus
- ☐ Belgium
- ☐ Bosnia and Herzegovina
- ☐ Bulgaria
- ☐ Croatia
- ☐ Cyprus
- ☐ Czechia
- ☐ Denmark
- ☐ Estonia
- ☐ Faroes
- ☐ Finland
- ☐ France
- ☐ Georgia
- ☐ Germany
- ☐ Gibraltar
- ☐ Greece
- ☐ Guernsey

## Covid-19 Testing EU/EEA & UK

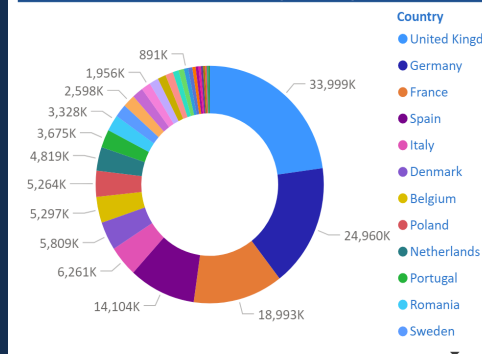
Country

All

Reported Date (From - To)

04/01/2020 07/11/2020

Tests done by Country



Tests done Vs Confirmed Cases



# Covid-19 Prediction/ Reporting

Microsoft AzureData Factorycovid-reporting-adf

Dashboard

Runs

Pipeline runs

Trigger runs

Runtimes & sessions

Integration runtimes

Data flow debug

Notifications

Alerts & metrics

Pipeline runs

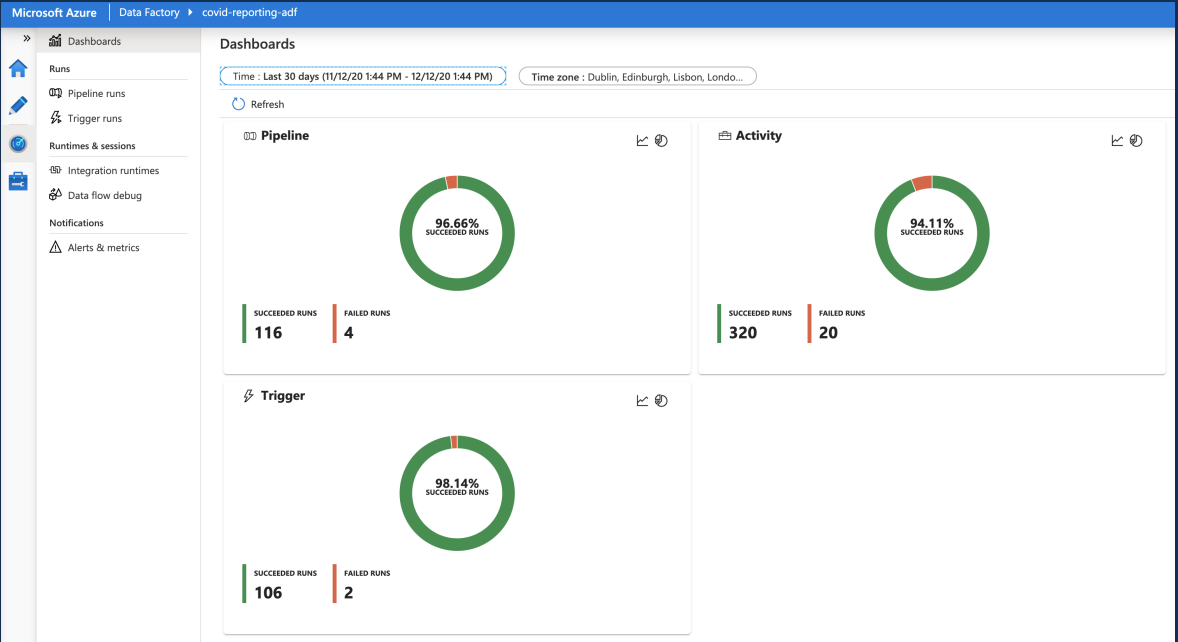
Triggered Debug Rerun Cancel Refresh

list Gantt

Search by run ID or name Dublin, Edinburgh, Li... Last 7 days Pipeline name: All Status: All Runs: Latest runs Add filter Copy filters

Showing 1 - 35 items

Pipeline name	Run start	Run end	Duration	Triggered by	Status	Run	Parameters	Annotations	Error	Run ID
pl_sqllze_hospital_admissions...	12/10/20, 12:08:18 AM	12/10/20, 12:08:26 AM	00:00:08	tr_sqllze_hospital_admissi...	Succeeded	Original				6c6daa32-5a16-4f8a-...
pl_sqllze_cases_and_deaths_d...	12/10/20, 12:08:05 AM	12/10/20, 12:08:18 AM	00:00:13	tr_sqllze_cases_and_deat...	Succeeded	Original				3d6e4217c-a018-4947-...
pl_process_hospital_admission...	12/10/20, 12:01:53 AM	12/10/20, 12:08:06 AM	00:06:13	tr_process_hospital_admi...	Succeeded	Original				a216226f-e230-4ac3-...
pl_process_cases_and_deaths...	12/10/20, 12:01:47 AM	12/10/20, 12:07:55 AM	00:06:07	tr_process_cases_and_de...	Succeeded	Original				38492eb6-383f-4afa-...
pl_ingest_ecdc_data	12/10/20, 12:00:12 AM	12/10/20, 12:01:30 AM	00:01:17	tr_ingest_ecdc_data	Succeeded	Original				d17a5375-1153-4ca2-...
pl_sqllze_cases_and_deaths_d...	12/9/20, 12:07:42 AM	12/9/20, 12:07:54 AM	00:00:12	tr_sqllze_cases_and_deat...	Succeeded	Original				3d6d1837-0d6f-4796-...
pl_sqllze_hospital_admissions...	12/9/20, 12:07:06 AM	12/9/20, 12:07:15 AM	00:00:08	tr_sqllze_hospital_admissi...	Succeeded	Original				919741a2-94d9-43fe-...
pl_process_cases_and_deaths...	12/9/20, 12:01:32 AM	12/9/20, 12:07:31 AM	00:06:09	tr_process_cases_and_de...	Succeeded	Original				1bfb4a26-12b5-4563-...
pl_process_hospital_admission...	12/9/20, 12:01:16 AM	12/9/20, 12:06:55 AM	00:05:39	tr_process_hospital_admi...	Succeeded	Original				0bdabaaf-1dea-42ef-...
pl_ingest_ecdc_data	12/9/20, 12:00:13 AM	12/9/20, 12:00:59 AM	00:00:45	tr_ingest_ecdc_data	Succeeded	Original				30daacfe-6e8a-497e-...
pl_sqllze_hospital_admissions...	12/8/20, 12:08:20 AM	12/8/20, 12:08:27 AM	00:00:07	tr_sqllze_hospital_admissi...	Succeeded	Original				77ab3327-4a45-406f-...
pl_sqllze_cases_and_deaths_d...	12/8/20, 12:07:55 AM	12/8/20, 12:08:08 AM	00:00:13	tr_sqllze_cases_and_deat...	Succeeded	Original				e77122d5-5ecd-4646-...
pl_process_cases_and_deaths...	12/8/20, 12:01:33 AM	12/8/20, 12:07:44 AM	00:06:10	tr_process_cases_and_de...	Succeeded	Original				0c9e0a3-a8e1-4337-...
pl_process_hospital_admission...	12/8/20, 12:01:27 AM	12/8/20, 12:08:10 AM	00:06:43	tr_process_hospital_admi...	Succeeded	Original				e44ca7d3-8cd2-4f7b-...
pl_ingest_ecdc_data	12/8/20, 12:00:12 AM	12/8/20, 12:01:10 AM	00:00:57	tr_ingest_ecdc_data	Succeeded	Original				521a9fc7-ba89-4660-...
pl_sqllze_cases_and_deaths_d...	12/7/20, 12:08:07 AM	12/7/20, 12:08:20 AM	00:00:13	tr_sqllze_cases_and_deat...	Succeeded	Original				8a3572c4-2206-4247-...
pl_sqllze_hospital_admissions...	12/7/20, 12:07:34 AM	12/7/20, 12:08:35 AM	00:01:01	tr_sqllze_hospital_admissi...	Succeeded	Original				4956c433-f7f3-427b-...
pl_process_cases_and_deaths...	12/7/20, 12:01:16 AM	12/7/20, 12:07:57 AM	00:06:40	tr_process_cases_and_de...	Succeeded	Original				69642f93-8bdc-4964-...
pl_process_hospital_admission...	12/7/20, 12:01:10 AM	12/7/20, 12:07:23 AM	00:06:13	tr_process_hospital_admi...	Succeeded	Original				9a08c76a-90bc-464d-...
pl_ingest_ecdc_data	12/7/20, 12:00:12 AM	12/7/20, 12:00:53 AM	00:00:40	tr_ingest_ecdc_data	Succeeded	Original				fb268d7d-2472-4bcc-...





# Who is this course for

University students

IT Developers from other disciplines

AWS/ GCP/ On-prem Data Engineers

Data Architects

Data Scientists

# Who is this course not for

Your main focus is not learning Azure Data Factory

You are not interested in hands-on learning approach

Your only focus is Azure Data Engineering Certification

# Pre-requisites

No prior knowledge assumed

cloud fundamentals would be beneficial, not necessary

Basic knowledge on SQL would be beneficial, not necessary

Azure Account

# Our Commitments

Ask Questions, I will answer 😊

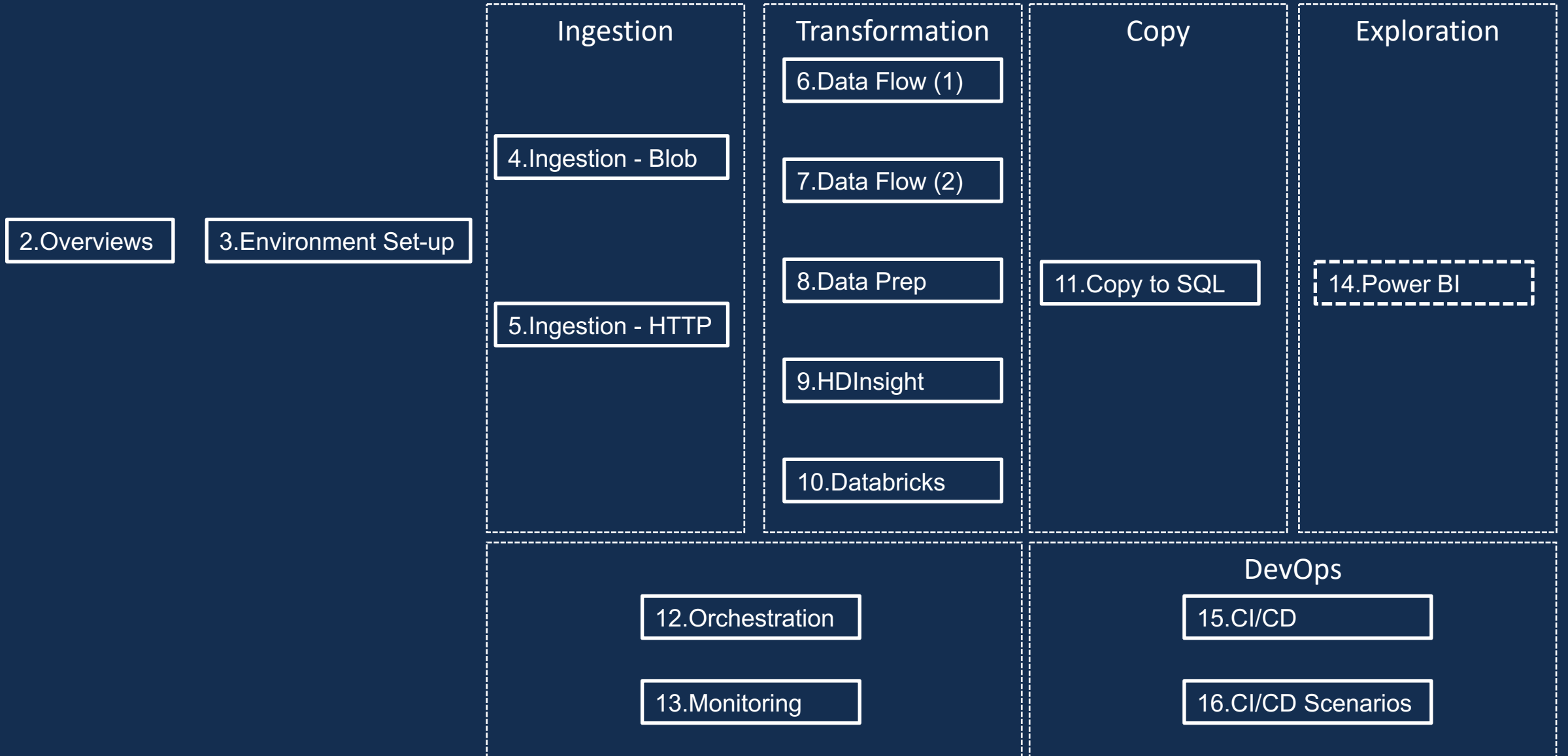
Keeping the course up to date

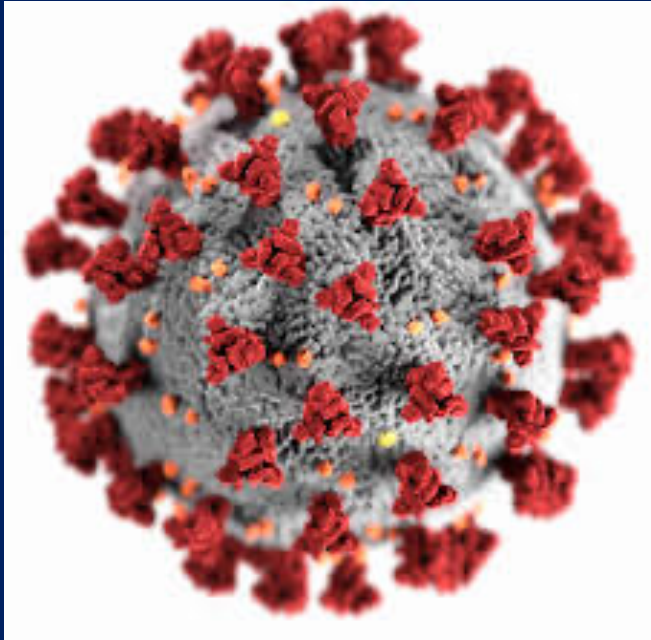
Udemy life time access

Udemy 30 day money back guarantee



# Course Structure

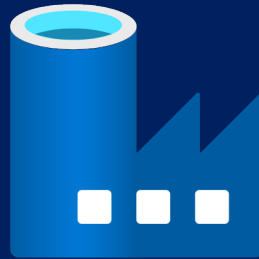




# Covid-19 Prediction/ Reporting

# Azure Data Factory Overview

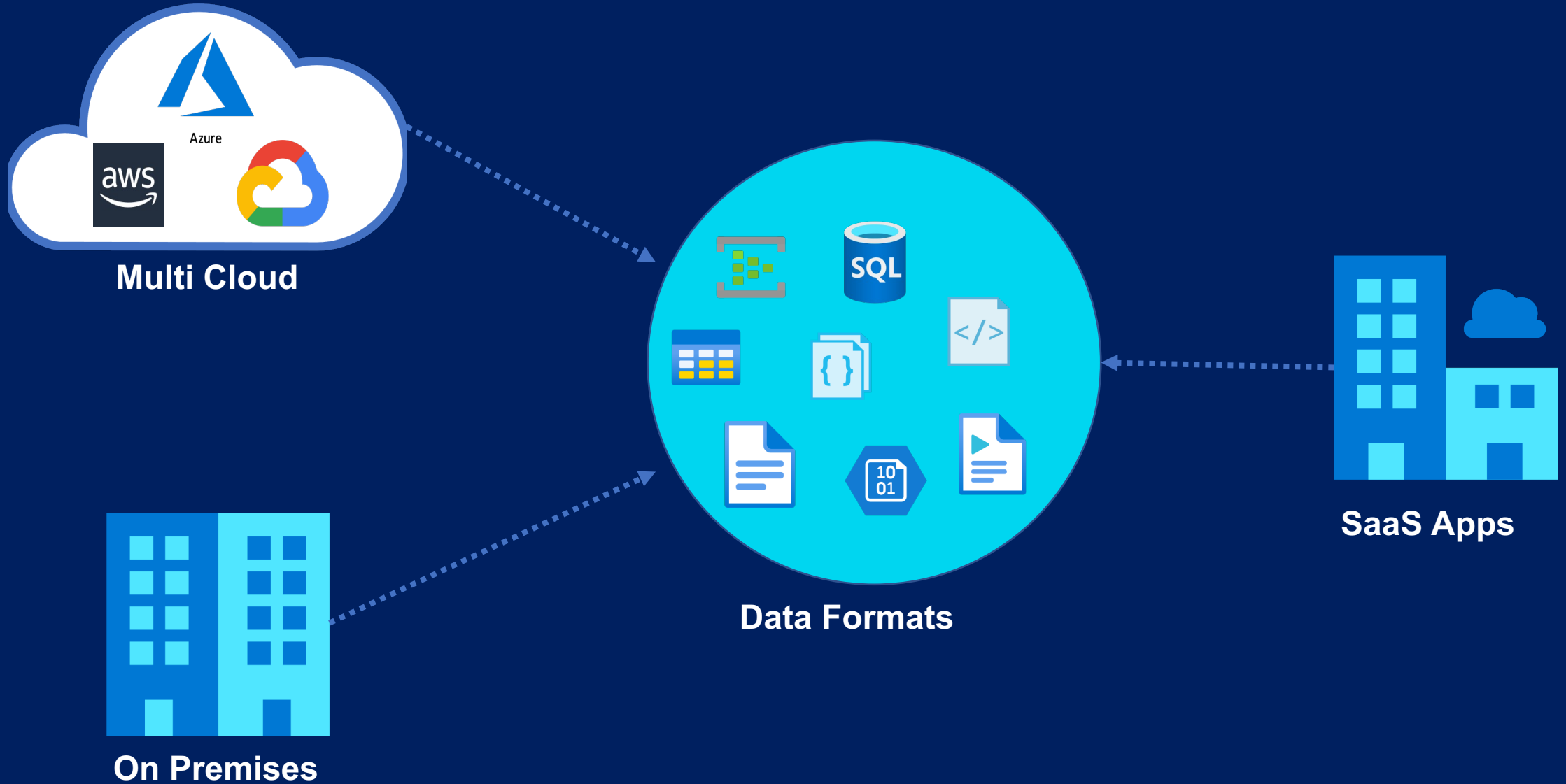
# What is Azure Data Factory



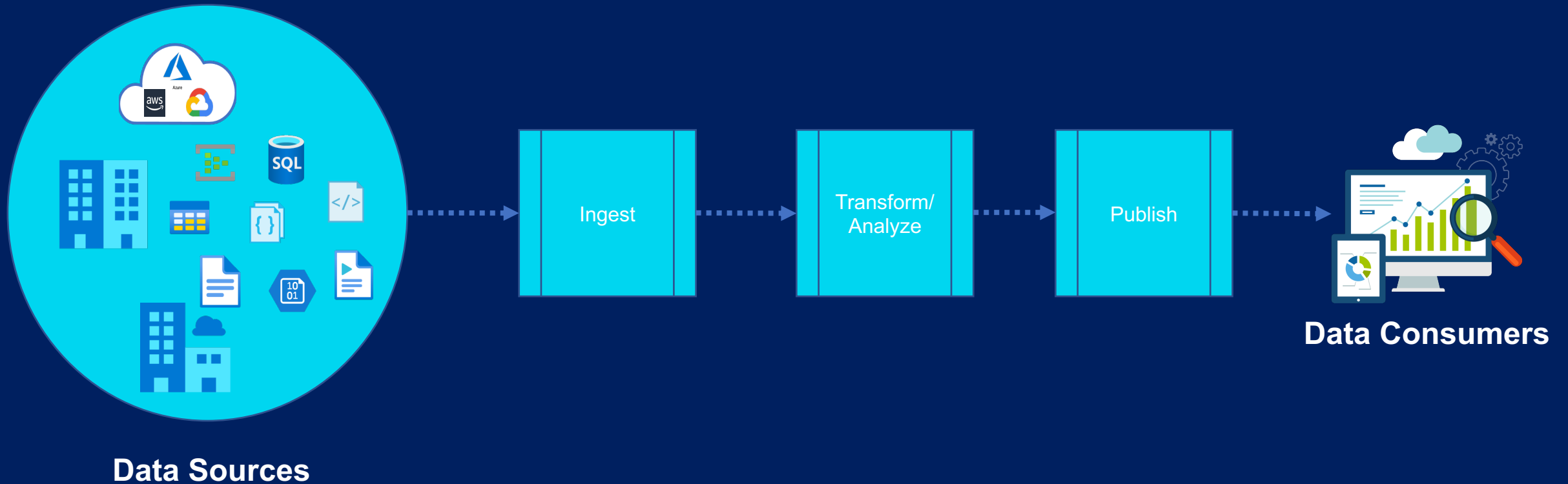
A fully managed, serverless data integration solution for ingesting, preparing and transforming all of your data at scale.



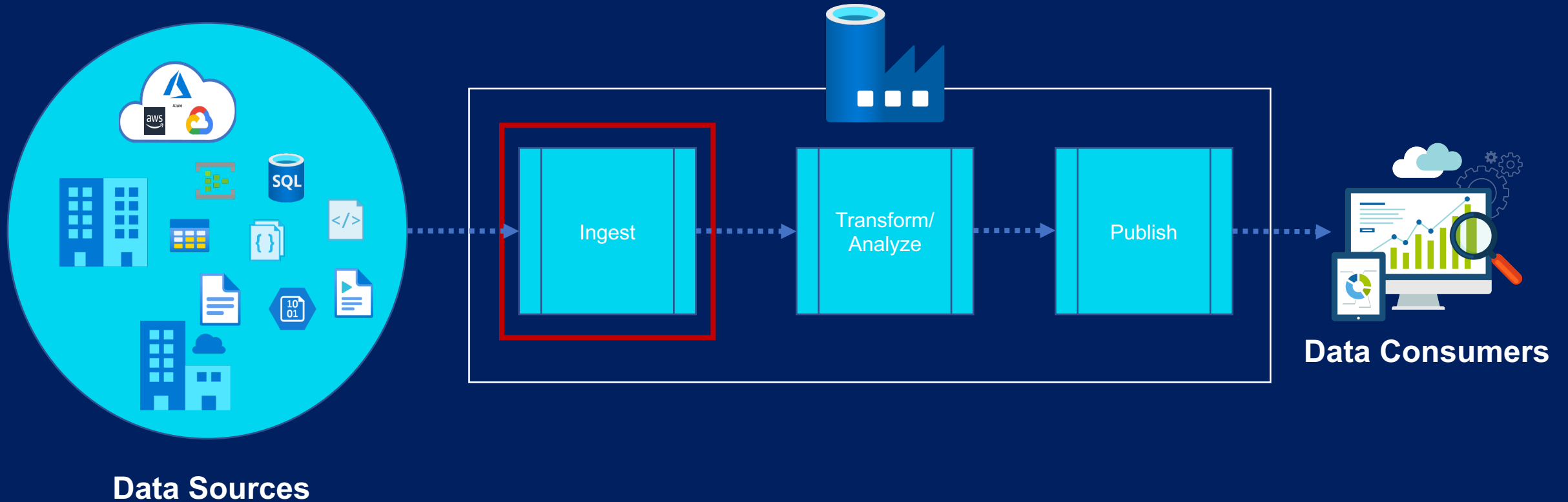
# The Data Problem



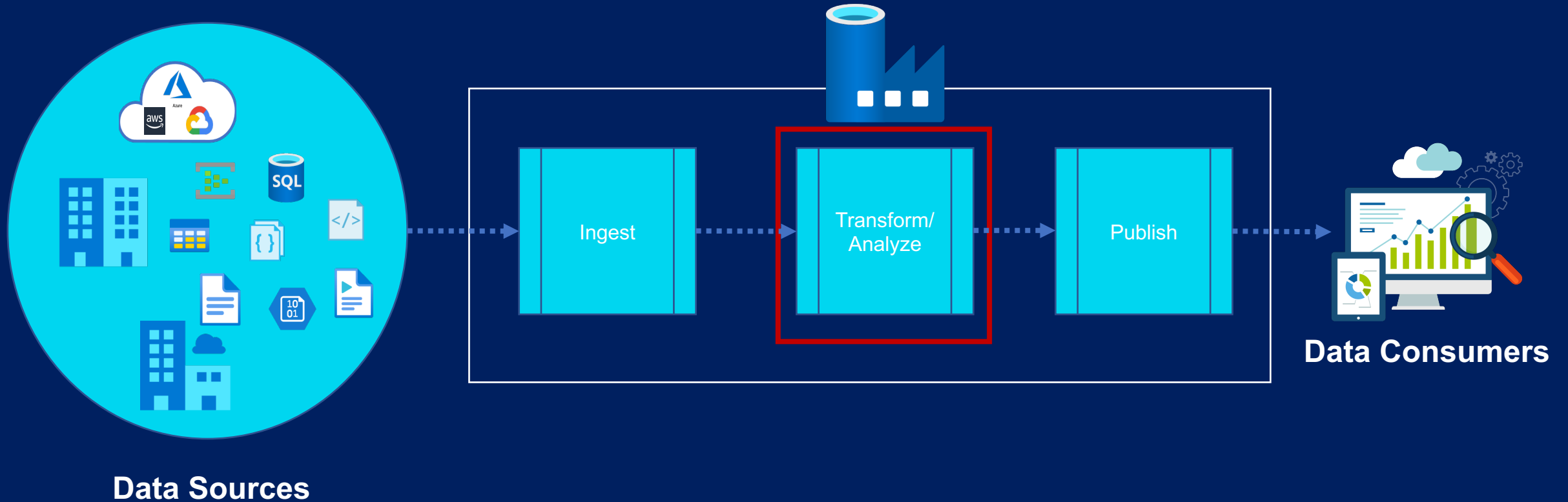
# The Data Problem



# The Data Problem

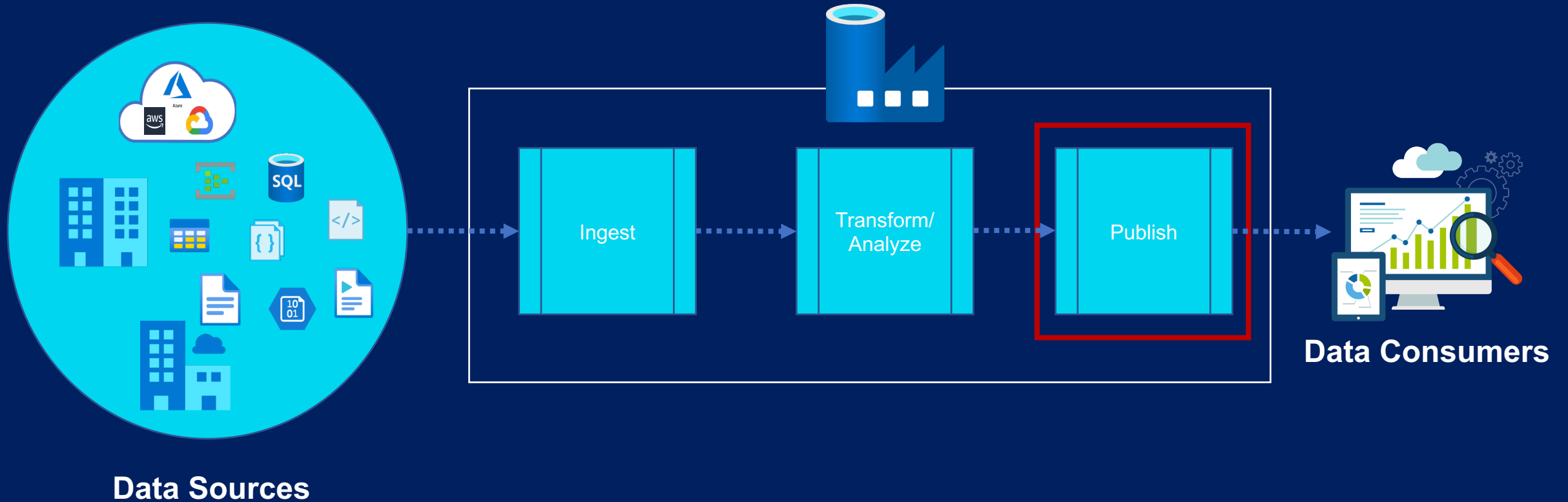


# The Data Problem

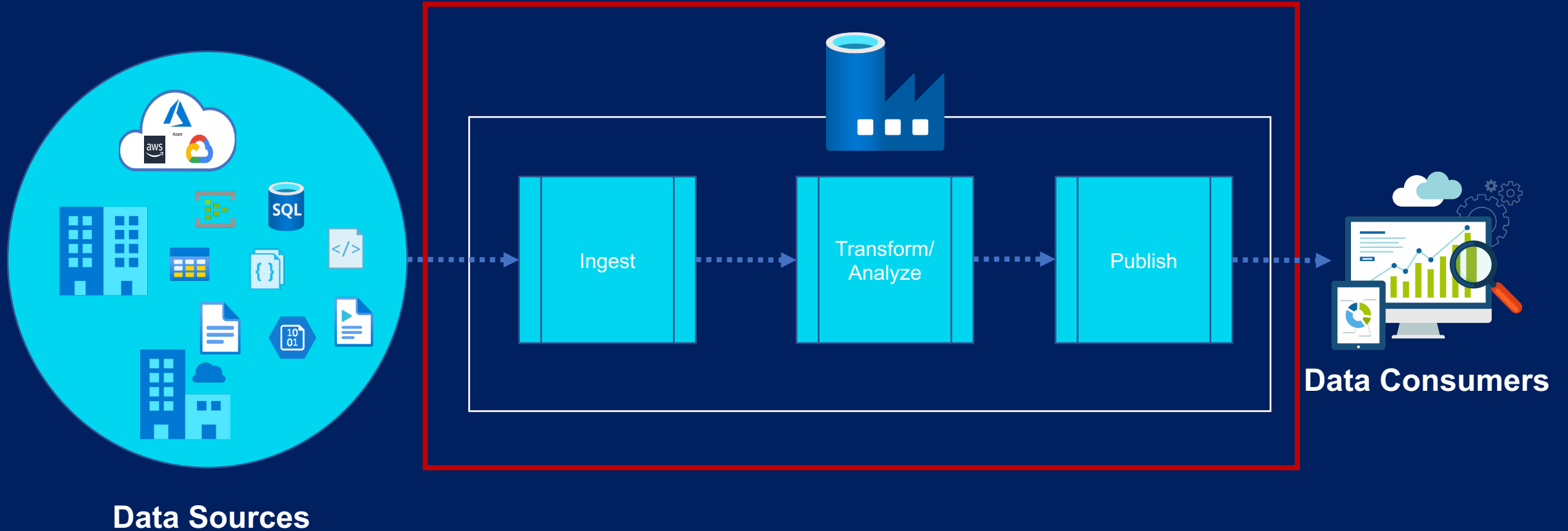




# The Data Problem



# The Data Problem



# What is Azure Data Factory



Fully Managed Service

Serverless

Data Integration Service

Data Transformation Service

Data Orchestration Service

*A fully managed, serverless data integration solution for ingesting, preparing and transforming all of your data at scale.*

# What Azure Data Factory Is Not



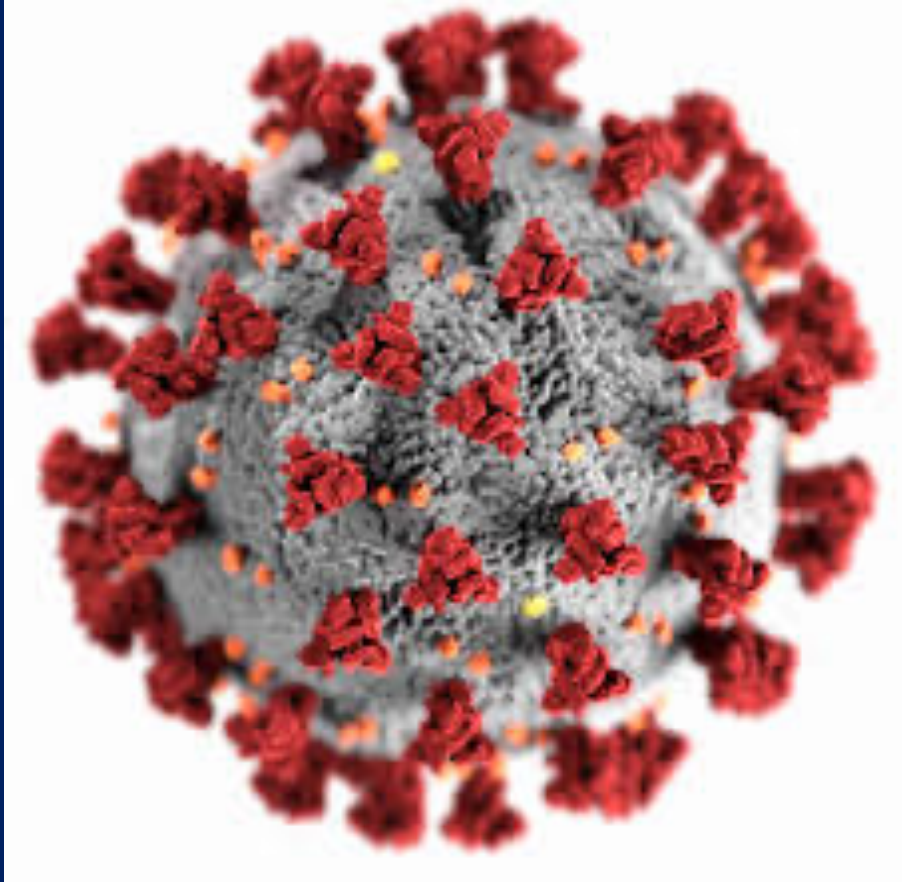
Data Migration Tool

Data Streaming Service

Suitable for Complex Data Transformations

Data Storage Service

# Project Overview



# Covid-19 Prediction/ Reporting

# Data Lake



Data Lake to be built with the following data to aid Data Scientists to predict the spread of the virus/ mortality

- Confirmed cases
- Mortality
- Hospitalization/ ICU Cases
- Testing Numbers
- Country's population by age group

# Data Warehouse



Data Warehouse to be built with the following data to aid Reporting on Trends

- Confirmed cases
- Mortality
- Hospitalization/ ICU Cases
- Testing Numbers



# Data Sources



## ECDC Website

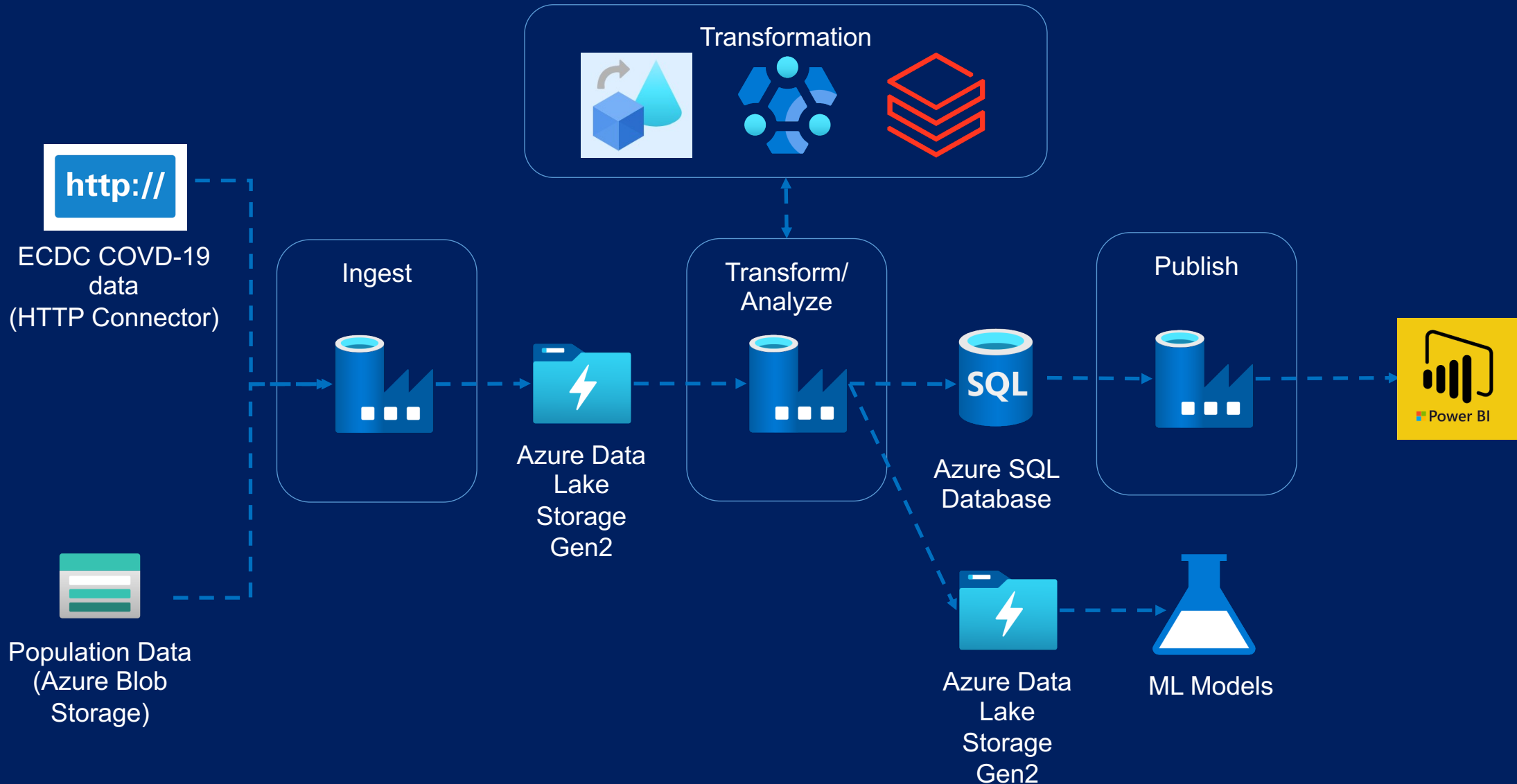
- Confirmed cases
- Mortality
- Hospitalization/ ICU Cases
- Testing Numbers

## Eurostat Website

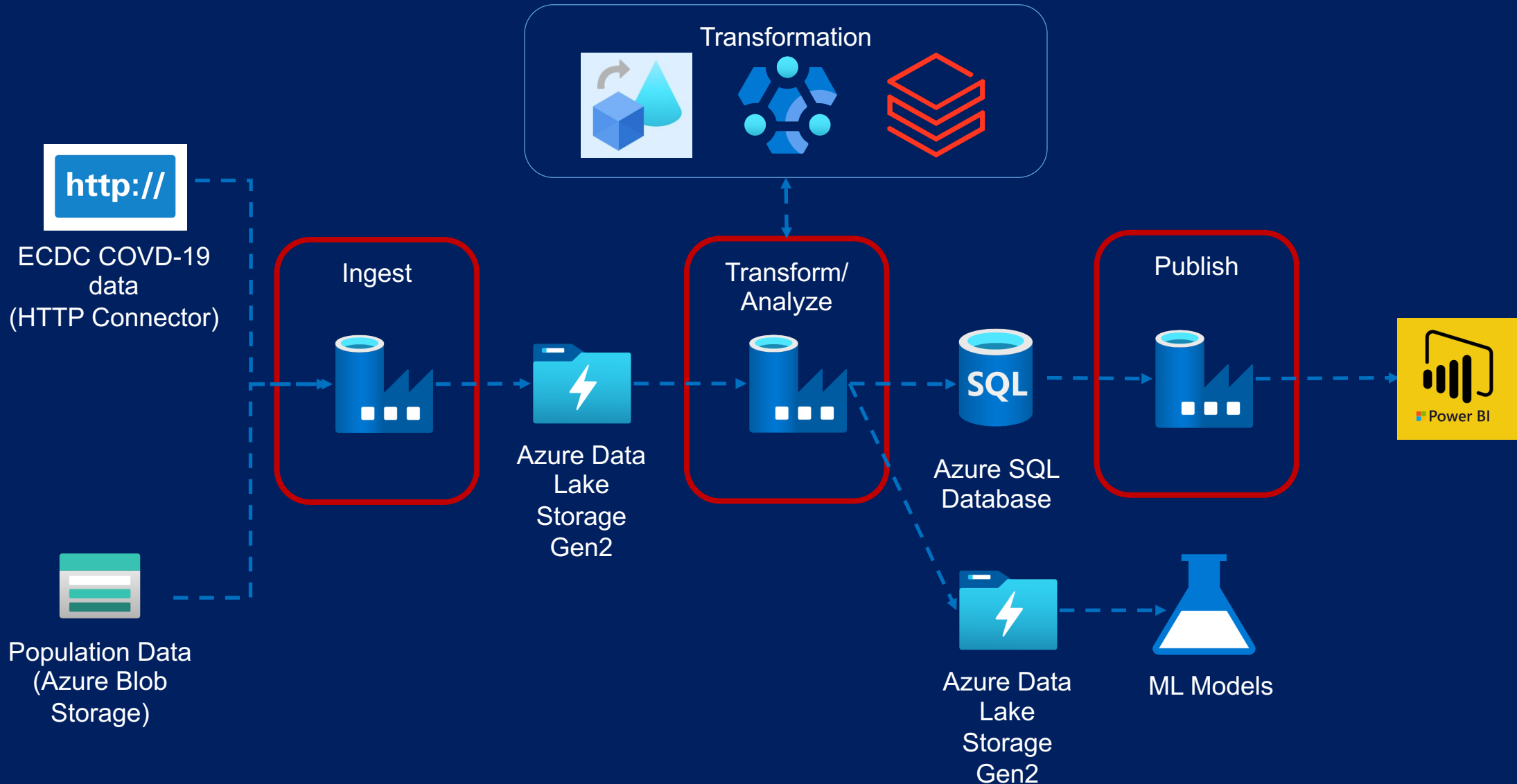
- Population by age

# Solution Architecture

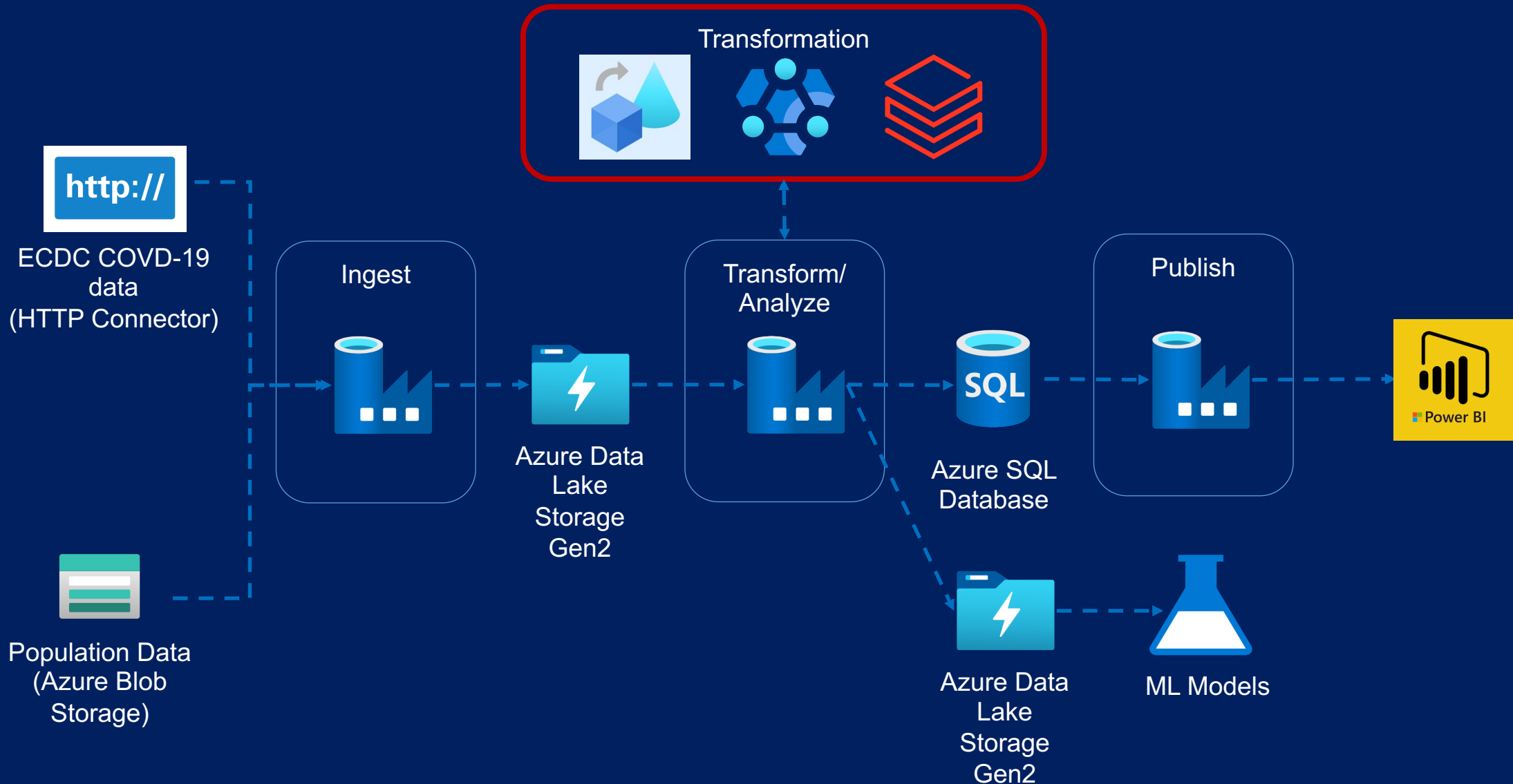
# Solution Architecture



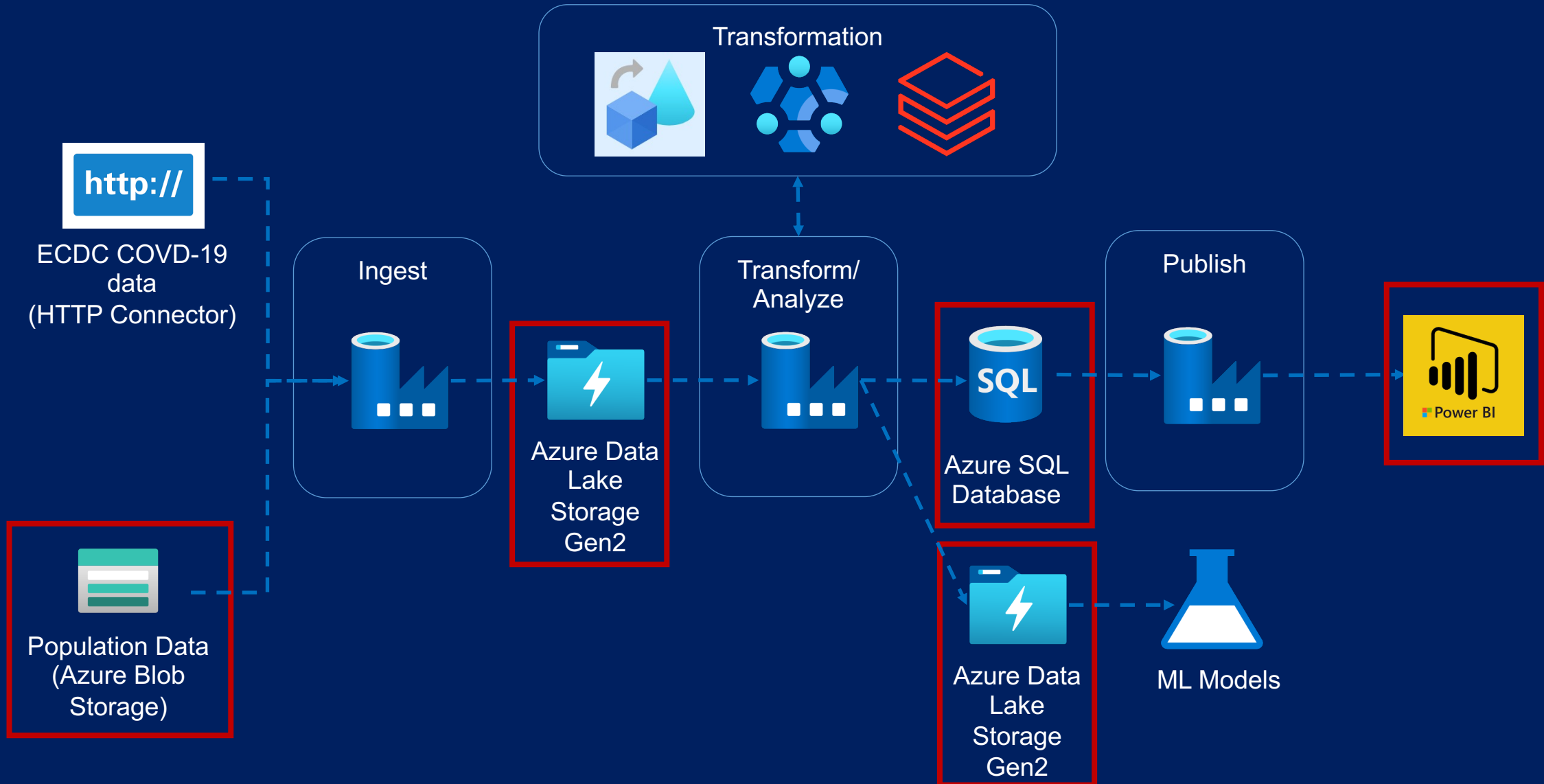
# Solution Architecture



# Solution Architecture



# Solution Architecture



# Storage Solutions

# Key Factors to Consider

## Structure of the data

Structured

Semi-Structured

Unstructured

## Operational needs

How often is the data accessed?

How quickly do we need to serve?

Need to run simple queries?

Need to run heavy analytical workload?

Accessed from multiple regions?



# Azure Databases



Azure SQL Database



Azure Database for MySQL



Azure Database for PostgreSQL



Azure Database for MariaDB



VM Images with Oracle, SQL Server etc.

# Azure Storage Account



Blob Storage



File Storage



Disk Storage

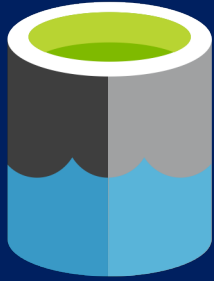


Table Storage



Queue Storage

# Azure Data Lake



Azure Data Lake Storage Gen2

Enhance Performance

Better Security

Enhance Management

# Azure Cosmos DB



Globally distributed

Multi Model

High Throughput

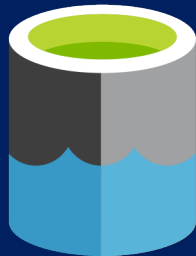
# Storage solutions used in this course



Azure SQL Database



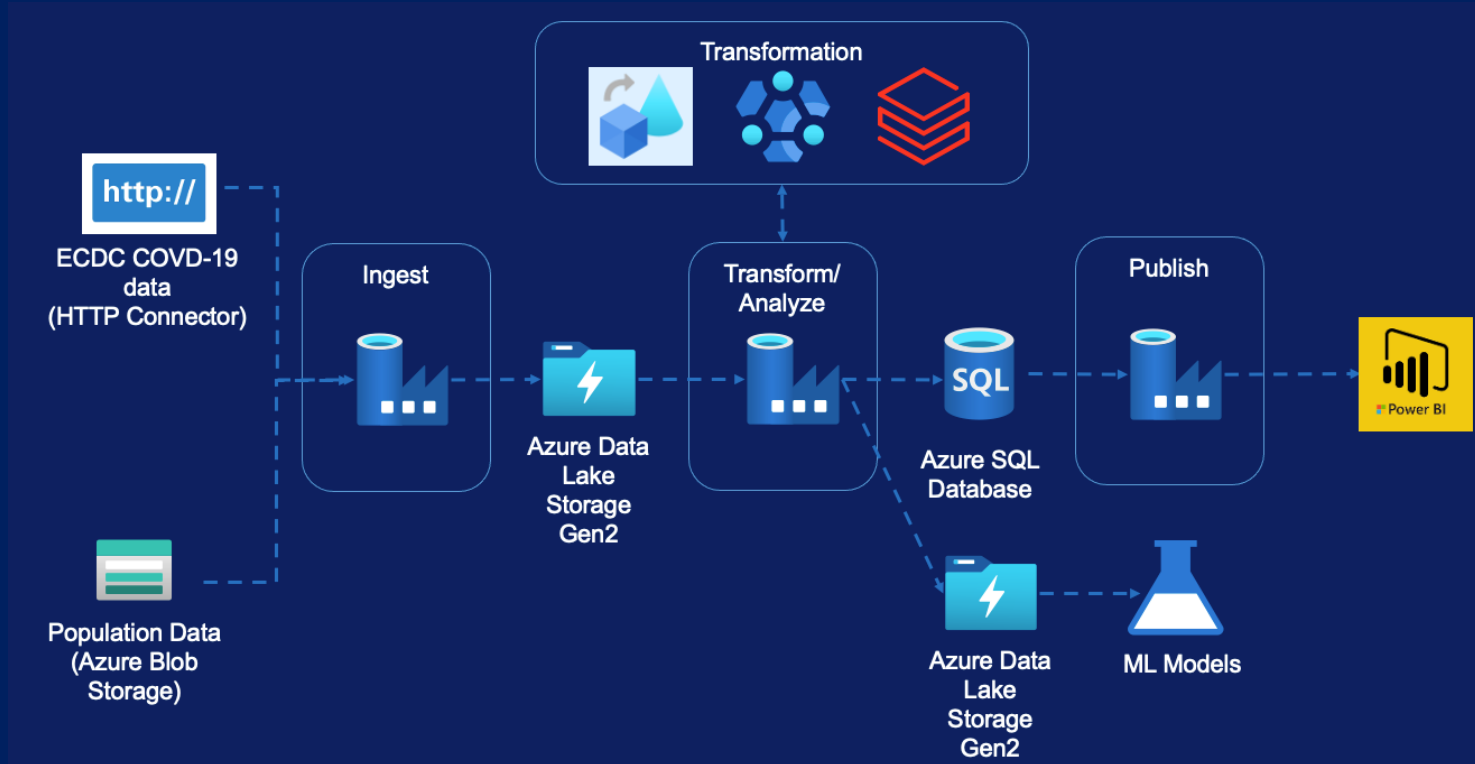
Azure Blob Storage



Azure Data Lake Storage Gen2

# Environment set-up

# Environment set-up



- Azure Subscription
- Data Factory
- Blob Storage Account
- Data Lake Storage Gen2
- Azure SQL Database
- Azure Databricks Cluster
- HD Insight Cluster

# Creating Azure Free Account





# Creating Azure Data Factory



# Creating Azure Storage Account



# Creating Azure Data Lake Gen2



# Creating Azure SQL Database

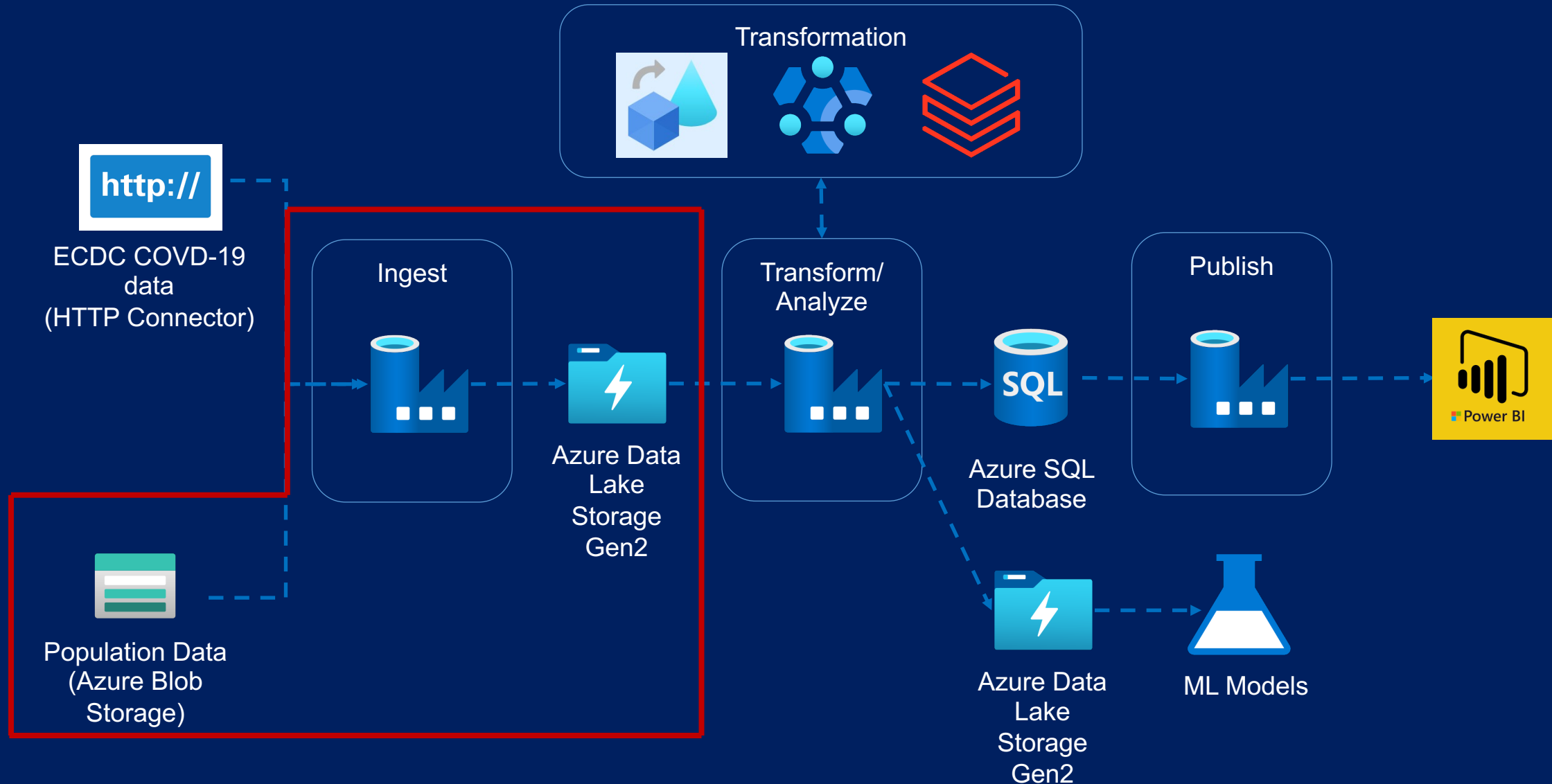


# Data Ingestion

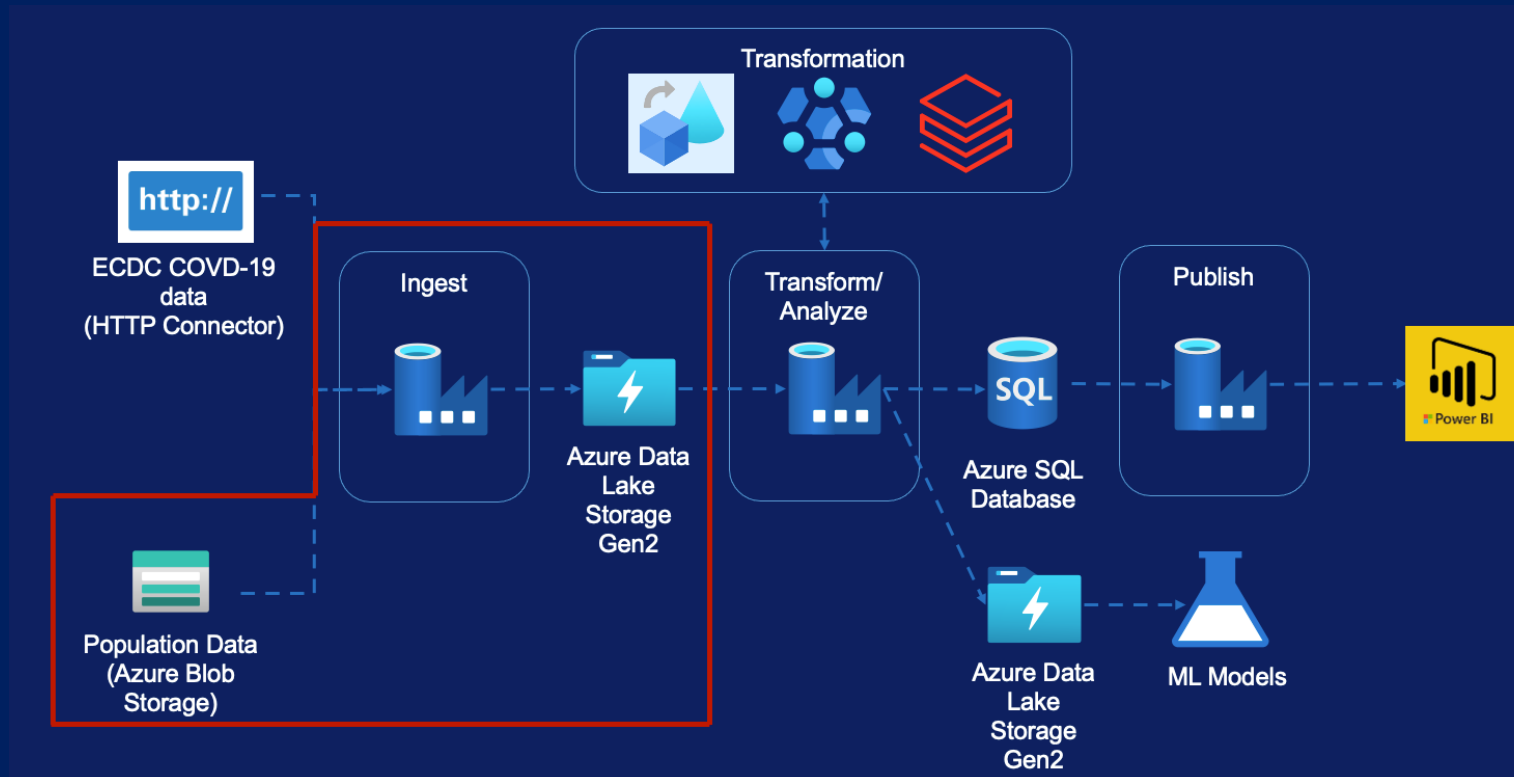
# Data Ingestion - Module Overview

## (Population by Age)

# Data Ingestion – Population Data



# Data Ingestion – Population Data



Copy Activity

Linked Services

Datasets

Pipeline

Validation Activity

If Condition Activity

Web Activity

Get Metadata Activity

Delete Activity

Trigger



## Copy Activity

Azure Blob Storage → Azure Data Lake

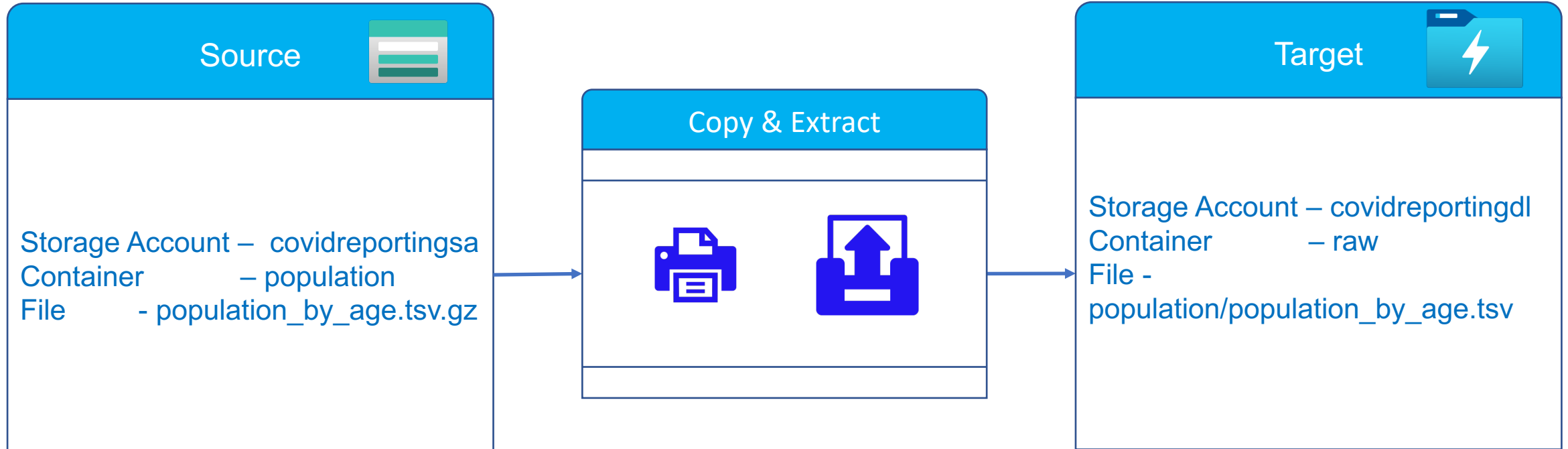
# Copy Activity



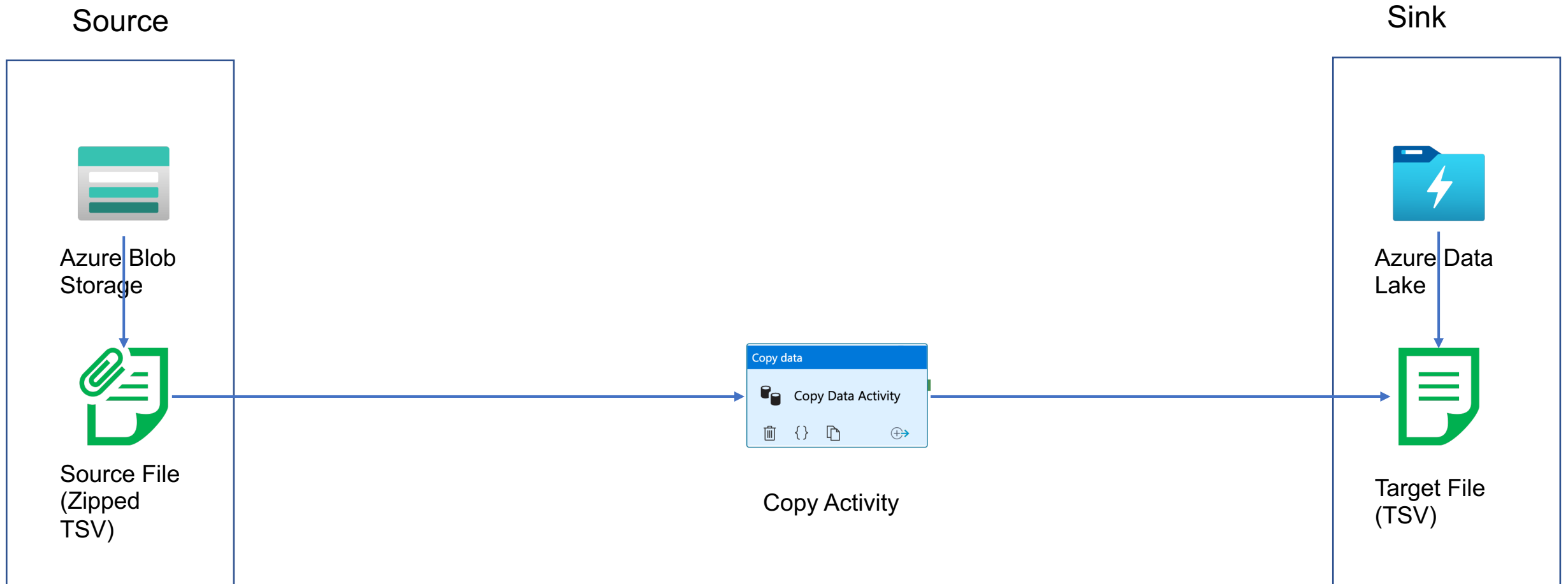
Ingest "population by age" for all EU Countries into the Data Lake to support the machine learning models to predict increase in Covid-19 mortality rates



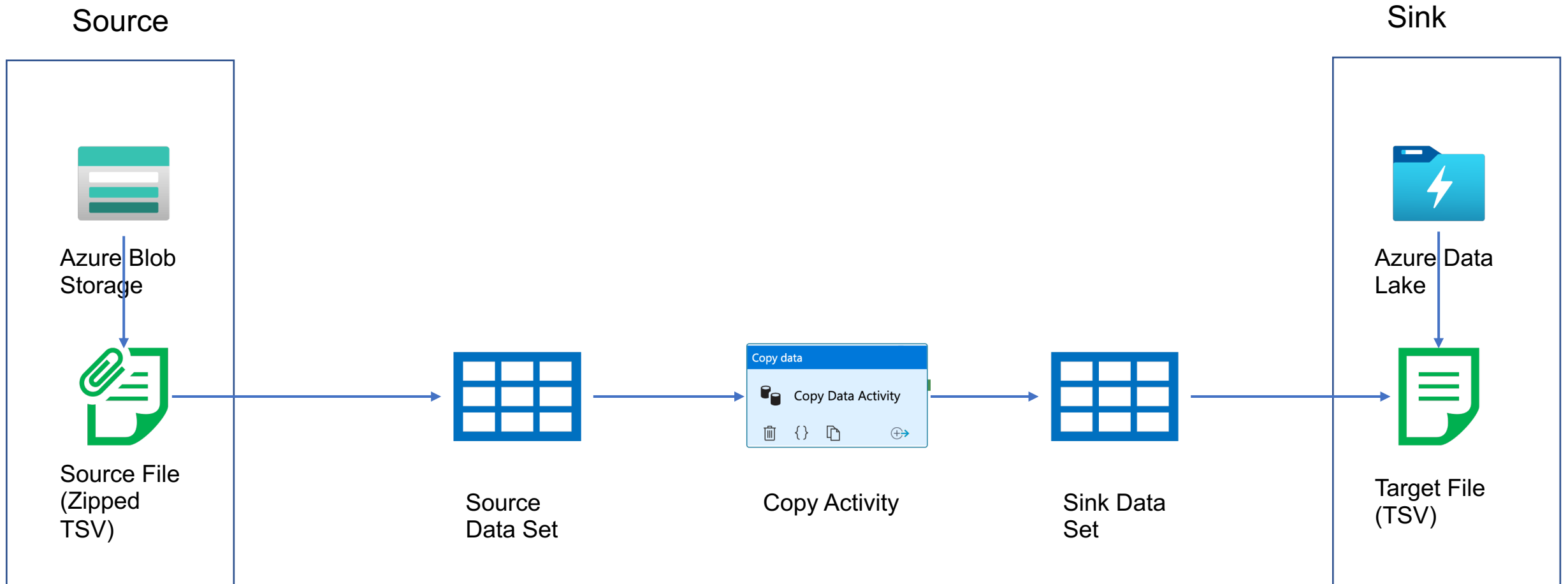
# Copy Activity



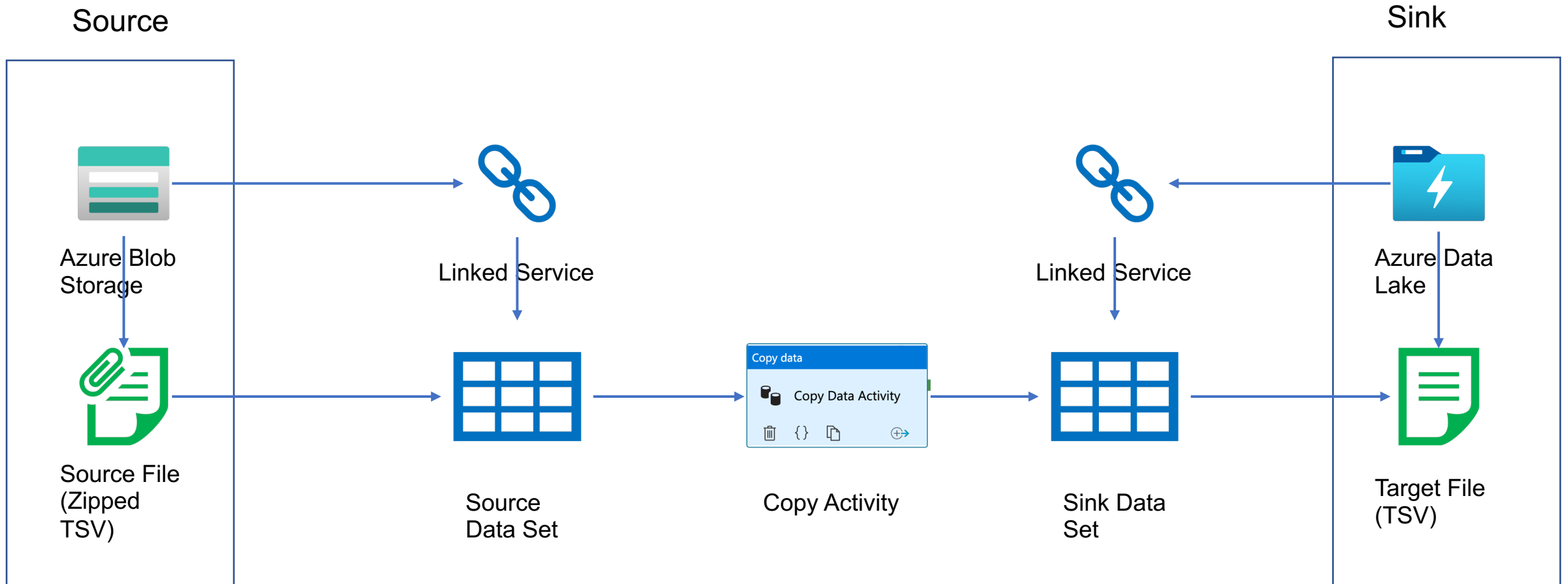
# Copy Activity



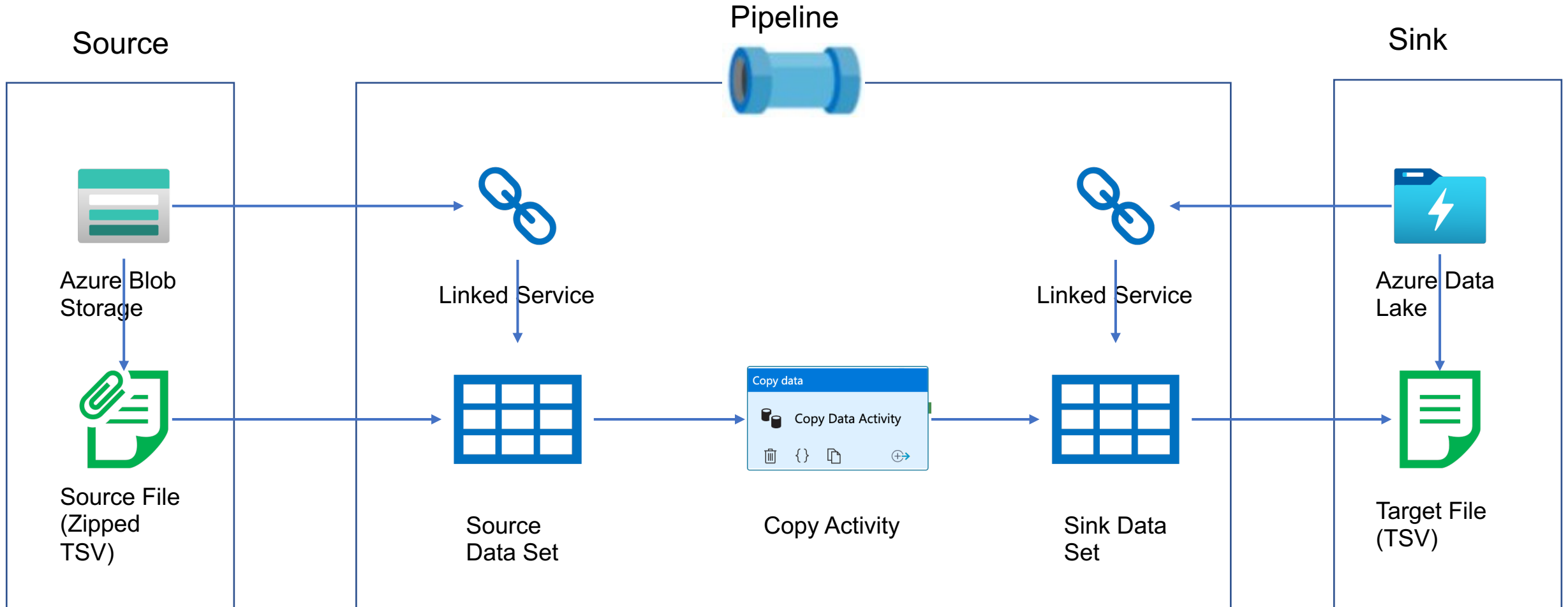
# Copy Activity



# Copy Activity



# Copy Activity

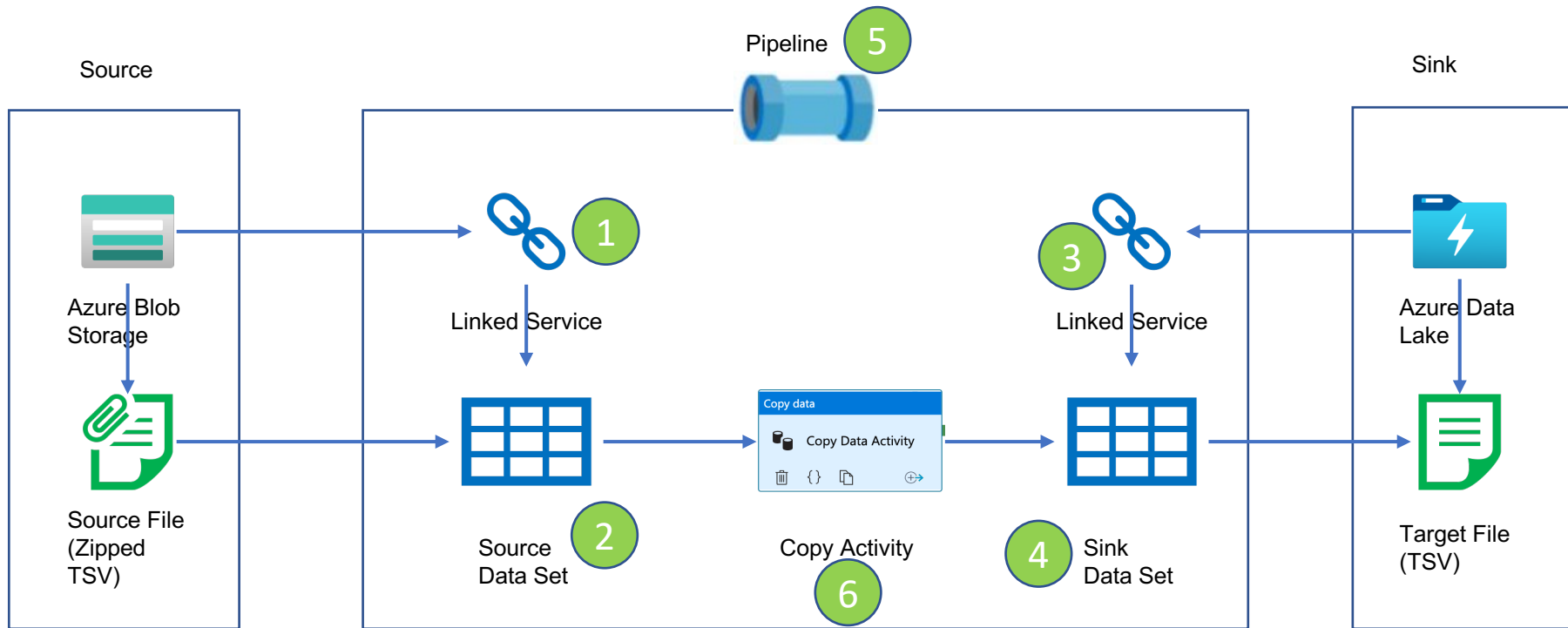


# Copy Activity From Azure Blob Storage





# Copy Activity



Storage Account: covidreportingsa  
Container: population  
File: population\_by\_age.tsv.gz

Storage Account: covidreportingdl  
Container: raw  
File: population/population\_by\_age.tsv

- 1 ls\_ablob\_covidreportingsa
- 2 ds\_population\_raw\_gz
- 3 ls\_adls\_covidreportingdl
- 4 ds\_population\_raw\_tsv
- 5 pl\_ingest\_population\_data
- 6 Copy Population Data

# Handling Real World Scenarios



# Scenario 1

Execute Copy Activity when the file becomes available



## Scenario 2

Execute Copy Activity only if file contents are as expected



# Scenario 3

Delete the source file on successful copy



# Scheduling Pipeline Execution





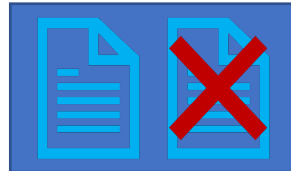
# Triggers



Schedule Trigger



Tumbling Window Trigger



Event Trigger



# Schedule Trigger



Runs on a calendar/ Clock



Supports periodic and specific times



Trigger to Pipeline is Many to Many



Can only be scheduled for a future time to start





# Tumbling Window Trigger



Runs at periodic intervals



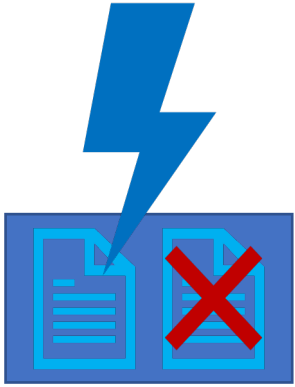
Windows are fixed sized, non-overlapping



Can be scheduled for the past windows/  
slices



Trigger to Pipeline is one to one



# Event Trigger



Runs in response to events



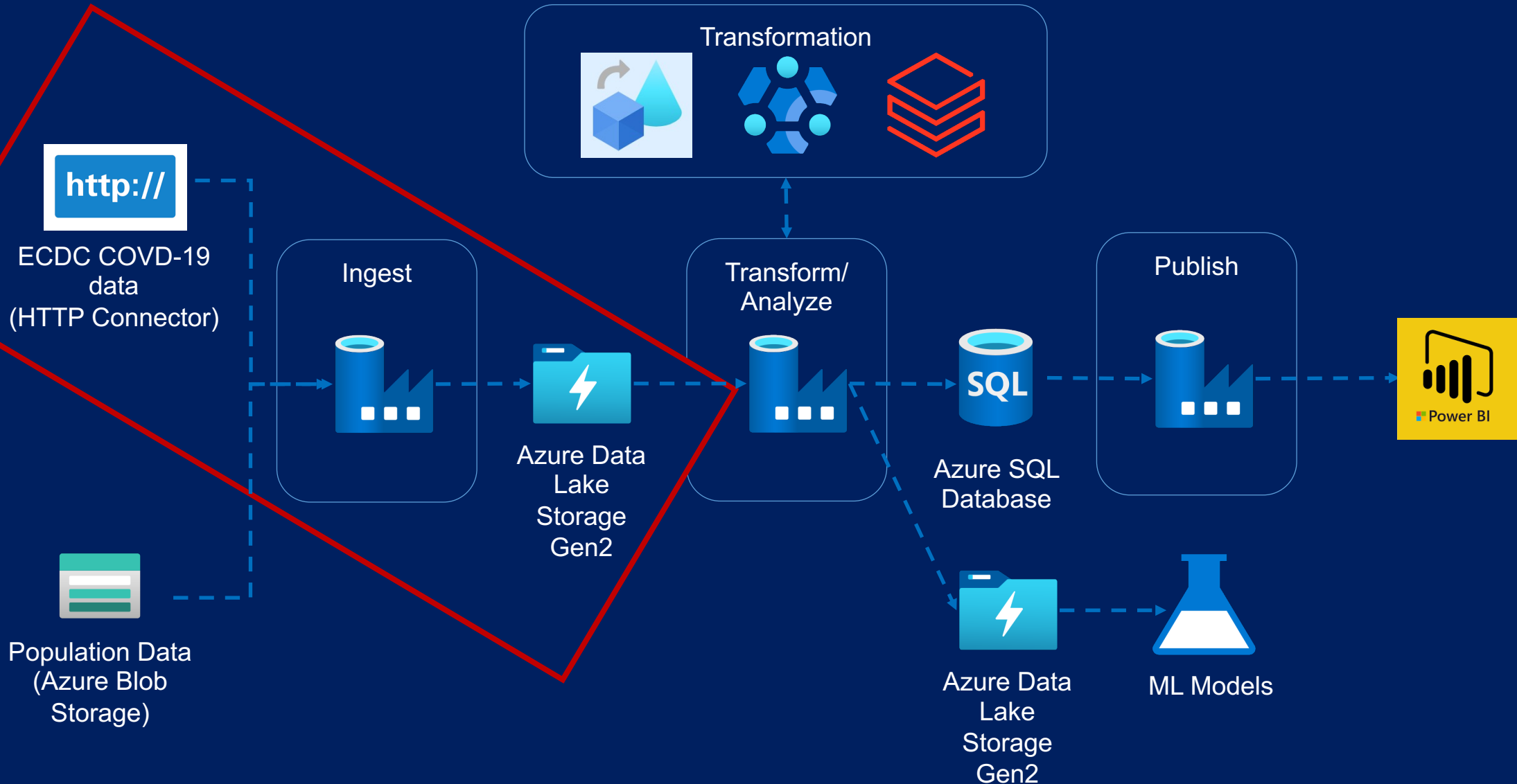
Events can be creation or deletion of Blobs/  
Files



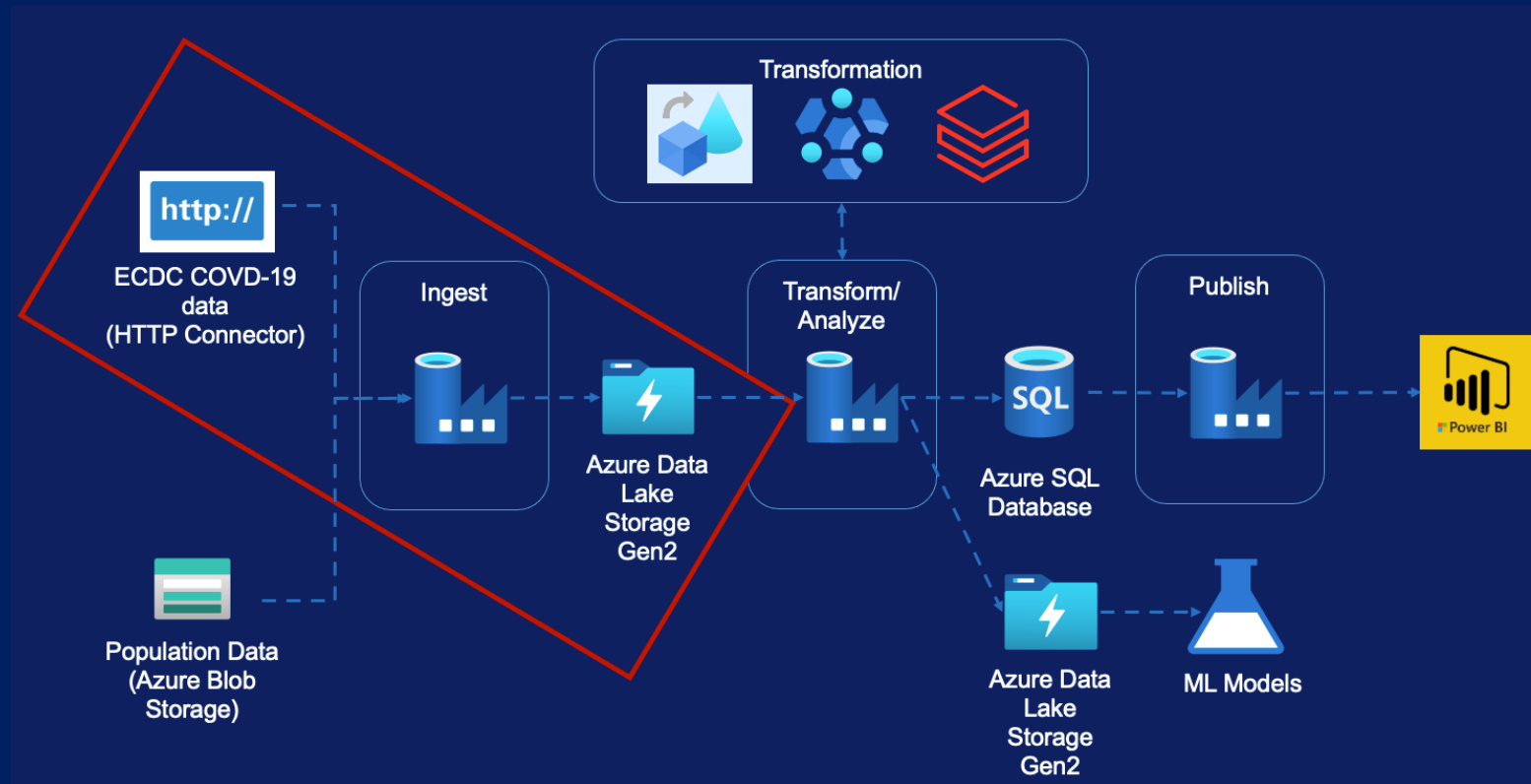
Trigger to Pipeline is Many to Many

# Data Ingestion - Module Overview (ECDC Data)

# Data Ingestion – ECDC Data



# Data Ingestion – ECDC Data



ECDC Data Overview

Create Initial Pipeline

Pipeline Variables

Pipeline Parameters

Lookup Activity

For Each Activity

Linked Service Parameters

Metadata driven pipeline

# Recent Changes to ECDC Data

# Recent Changes to ECDC Data

## Download COVID-19 datasets



ECDC switched to a weekly reporting schedule for the COVID-19 situation worldwide and in the EU/EEA and the UK on 17 December 2020. Hence, all daily updates have been discontinued from 14 December. ECDC will publish updates on the number of cases and deaths reported worldwide and aggregated by week every Thursday. The weekly data will be available as downloadable files in the following formats: XLSX, CSV, JSON and XML. As an exception, the weekly updates for the end-of-year festive season will be published on 23 December and 30 December 2020.

With the switch from daily to weekly reporting, ECDC will shift its Epidemic Intelligence (EI) resources from case counting to signal/event detection and resume its regular EI activities, which will include COVID-19 signal and event detection and analysis but also other potential threats.

- Granularity of the data changed from daily to weekly

- File structure is also different as a result

- Use GIT Repo - <https://github.com/cloudboxacademy/covid19>

# Data Ingestion

HTTP



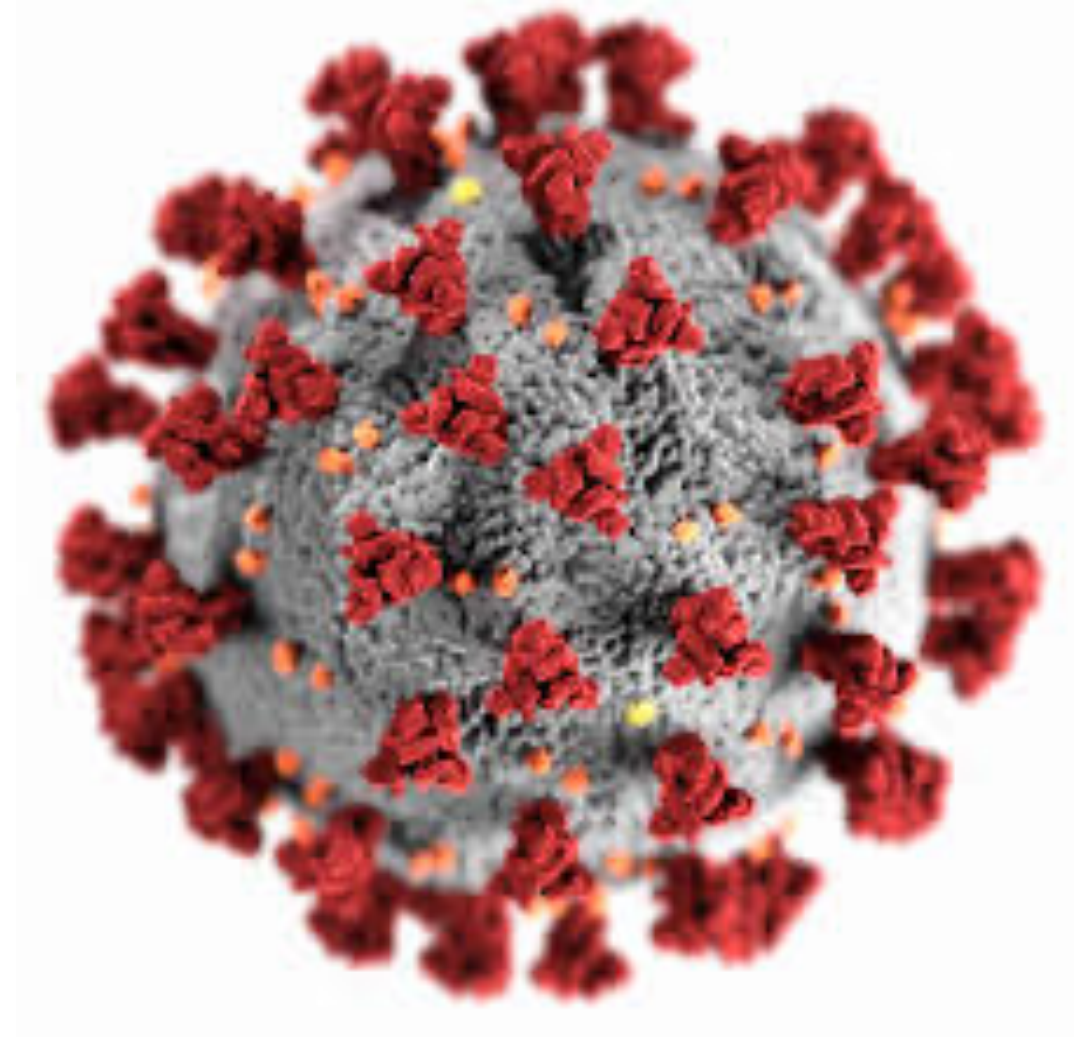
Azure Data Lake



# Data Ingestion Requirements

- Covid-19 new cases and deaths by Country
- Covid-19 Hospital admissions & ICU cases
- Covid-19 Testing Numbers
- Country Response to Covid-19

*URL - <https://www.ecdc.europa.eu/en/covid-19/data>*

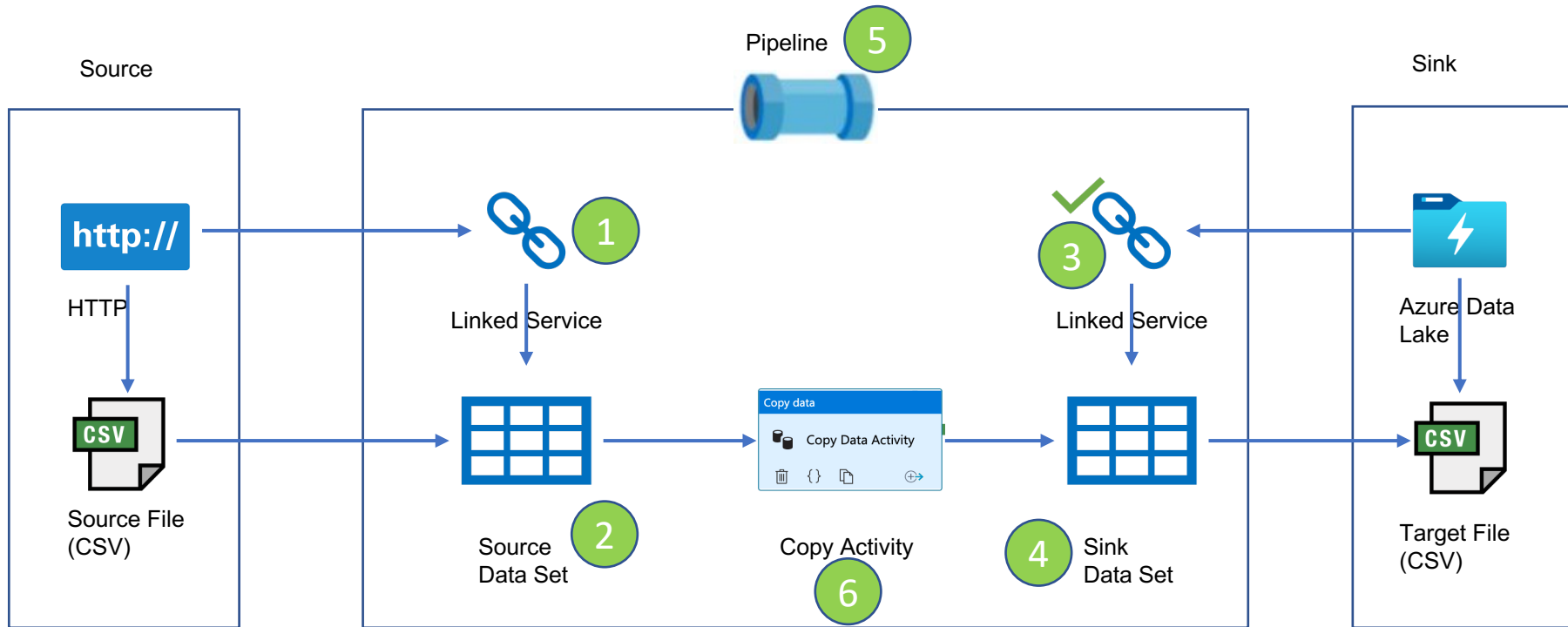


# Data Ingestion

## Case & Deaths Data

*URL - <https://www.ecdc.europa.eu/en/publications-data/data-national-14-day-notification-rate-covid-19>*

# Copy Activity – Case & Deaths Data

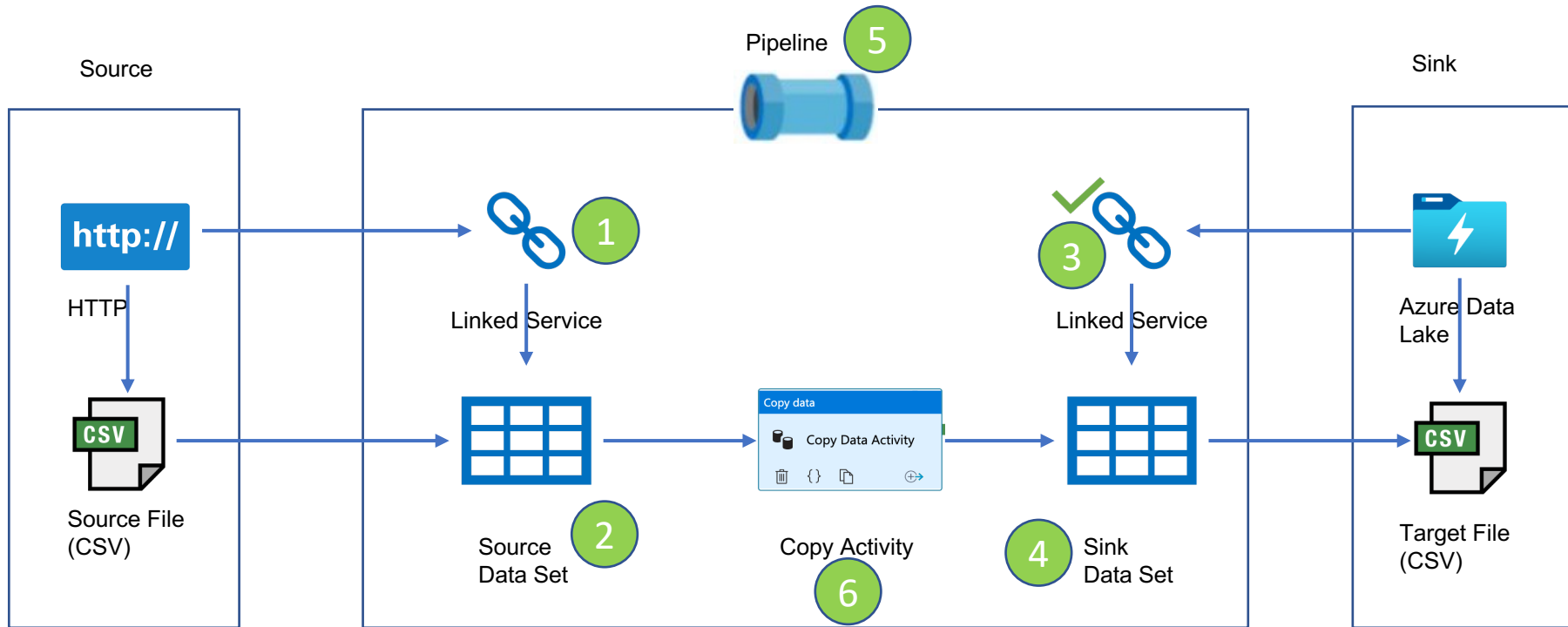


- 1 ls\_http\_opendata\_ecdc\_europa\_eu
- 2 ds\_cases\_deaths\_raw\_csv\_http
- 3 ls\_adls\_covidreportingdl ✓
- 4 ds\_cases\_deaths\_raw\_csv\_dl
- 5 pl\_ingest\_cases\_deaths\_data
- 6 Copy Cases And Deaths Data

URL:  
<https://opendata.ecdc.europa.eu/covid19/nationalcasedeath/csv>

Storage Account: covidreportingdl  
Container: raw  
File: ecdc/cases\_deaths.csv

# Copy Activity – Case & Deaths Data

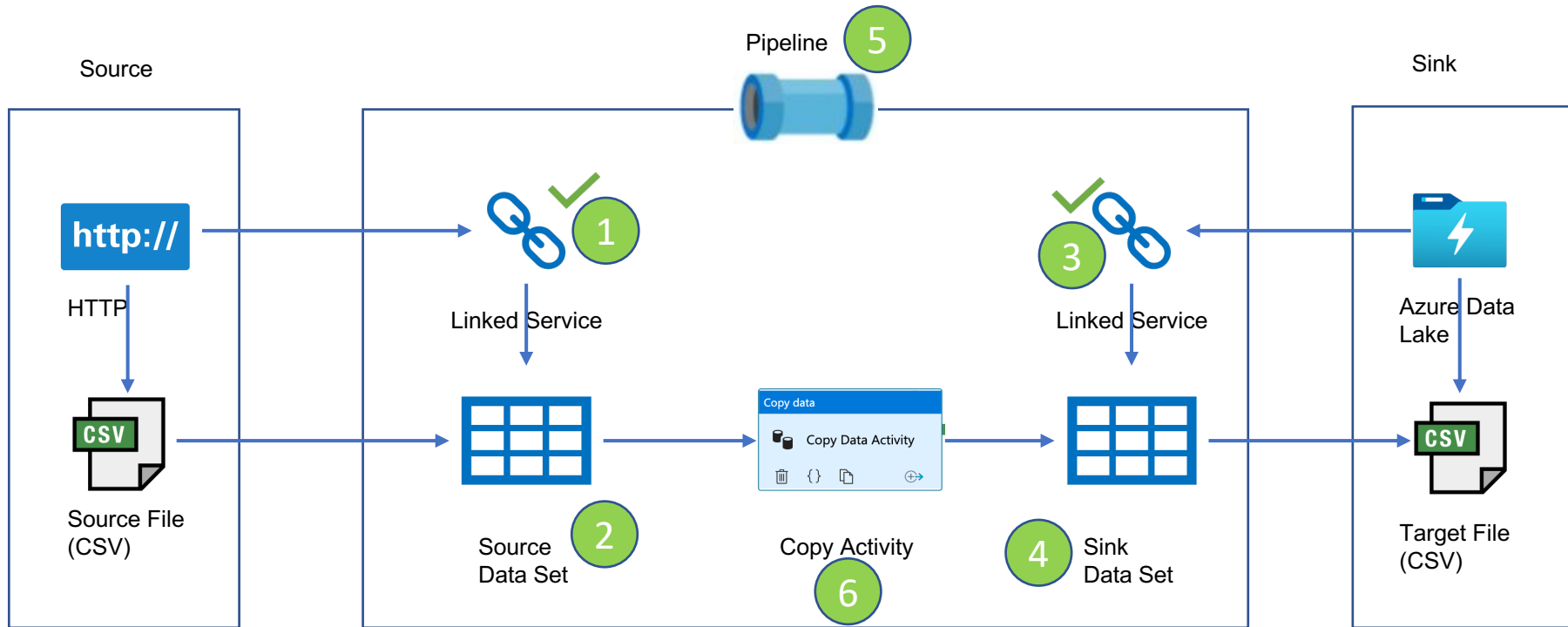


- 1 ls\_http\_opendata\_ecdc\_europa\_eu
- 2 ds\_cases\_deaths\_raw\_csv\_http
- 3 ls\_adls\_covidreportingdl ✓
- 4 ds\_cases\_deaths\_raw\_csv\_dl
- 5 pl\_ingest\_cases\_deaths\_data
- 6 Copy Cases And Deaths Data

URL:  
<https://opendata.ecdc.europa.eu/covid19/nationalcasedeath/csv>

Storage Account: covidreportingdl  
Container: raw  
File: ecdc/cases\_deaths.csv

# Copy Activity – Hospital Admission Data



- 1 ls\_http\_opendata\_ecdc\_europa\_eu ✓
- 2 ds\_hospital\_admissions\_raw\_csv\_http
- 3 ls\_adls\_covidreportingdl ✓
- 4 ds\_hospital\_admissions\_raw\_csv\_dl
- 5 pl\_ingest\_hospital\_admissions\_data
- 6 Copy Hospital Admissions Data

URL:  
<https://opendata.ecdc.europa.eu/covid19/hospitalicuadmissionrates/csv/data.csv>

Storage Account: covidreportingdl  
Container: raw  
File: ecdc/hospital\_admissions.csv

# Parameters & Variables

**Parameters** are external values passed into pipelines, datasets or linked services. The value cannot be changed inside a pipeline.

**Variables** are internal values set inside a pipeline. The value can be changed inside the pipeline using Set Variable or Append Variable Activity

# Differences

## Source

<https://opendata.ecdc.europa.eu/covid19/nationalcasedeath/csv>

<https://opendata.ecdc.europa.eu/covid19/hospitalicuadmissionrates/csv/data.csv>

<https://opendata.ecdc.europa.eu/covid19/testing/csv>

[https://www.ecdc.europa.eu/sites/default/files/documents/data\\_response\\_graphs\\_0.csv](https://www.ecdc.europa.eu/sites/default/files/documents/data_response_graphs_0.csv)

## Sink

raw/ecdc/case\_distribution.csv

raw/ecdc/hospital\_admission.csv

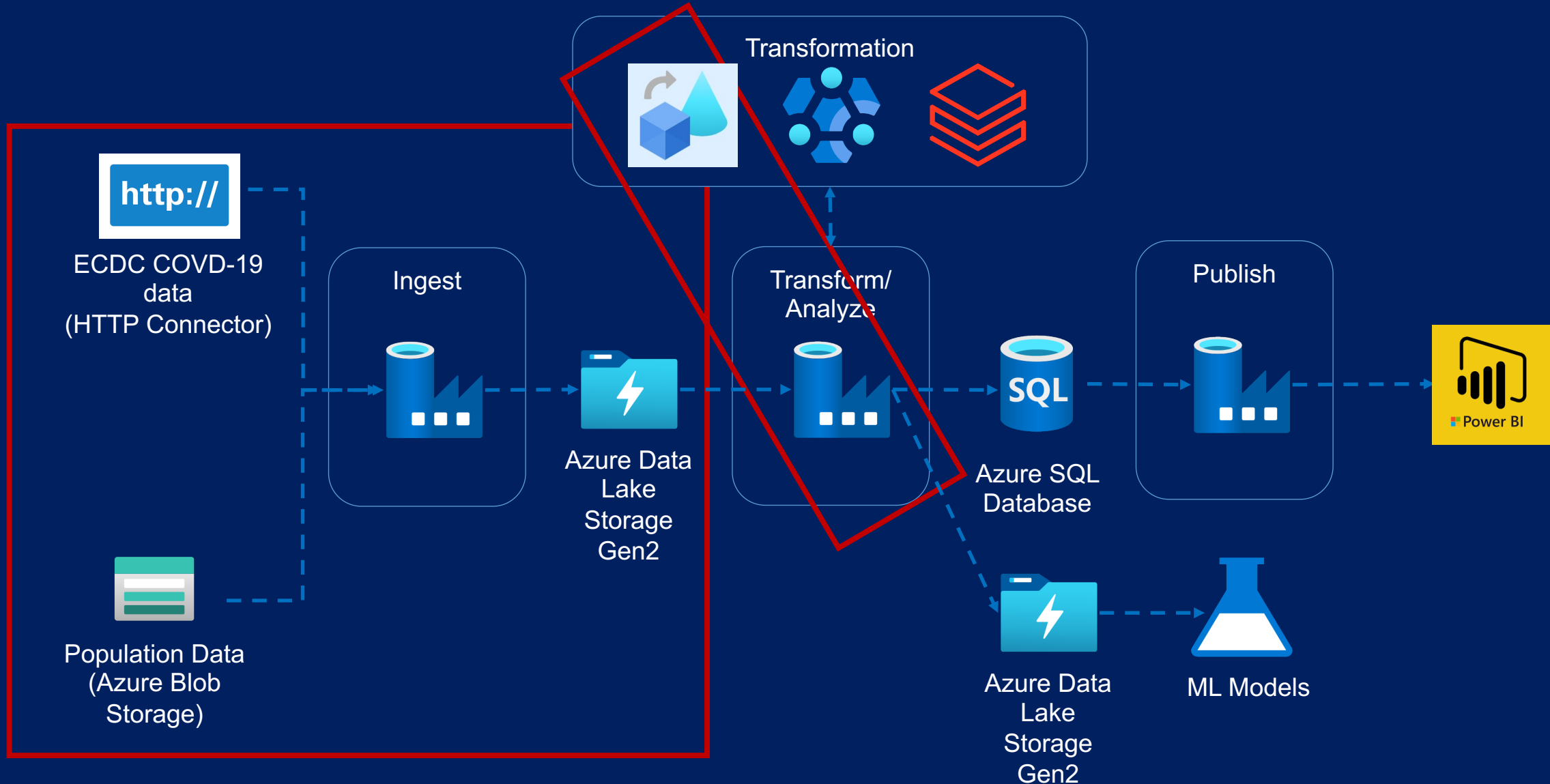
raw/ecdc/testing.csv

raw/ecdc/country\_response.csv

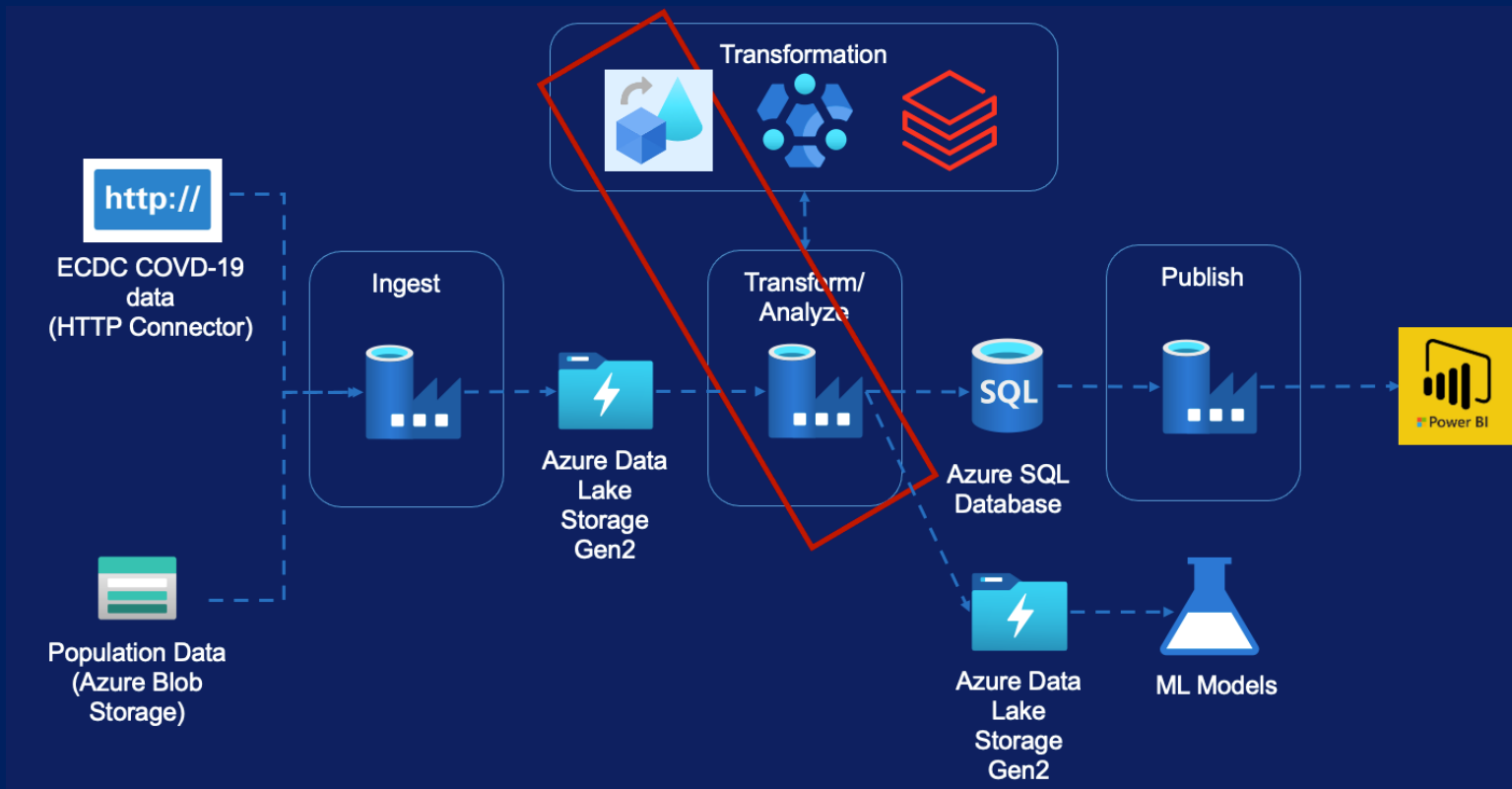
# Data Flows (1) - Module Overview (Cases & Deaths File)



# Data Flow – Cases & Deaths Data



# Data Flow – Cases & Deaths Data



Data Flow Overview

Requirement

Source Transformation

Filter Transformation

Select Transformation

Pivot Transformation

Lookup Transformation

Sink Transformation

Create Pipeline

# Data Flows

# Data Flows

## Features

- Code free data transformations
- Executed on Data Factory managed Databricks Spark clusters
- Benefits from Data factory scheduling and monitoring capabilities.

# Data Flows

## Types



### **Data flow**

Code free data transformation at scale



### **Wrangling Data Flow (Preview)**

Code free data preparation at scale

# Data Flows

## Limitations

- Only available in some regions

<https://docs.microsoft.com/en-us/azure/data-factory/concepts-data-flow-overview#available-regions>

- Limited set of connectors available

<https://docs.microsoft.com/en-us/azure/data-factory/data-flow-source#supported-sources>

- Not suitable for very complex logic

# Data Flows



# Transform Cases & Deaths Data





# Transform Cases & Deaths Data

## Raw File from ECDC

Column Name
country
country_code
<del>continent</del>
population
indicator
daily_count
date
<del>rate_14_day</del>
source

Europe  
Only

## Transformed File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit
population
cases_count
deaths_count
reported_date
source

# Transform Cases & Deaths Data

## Raw File from ECDC

Column Name
country
country_code
<del>continent</del> ✓
population
indicator
daily_count
date
<del>rate_14_day</del> ✓
source

Europe  
Only

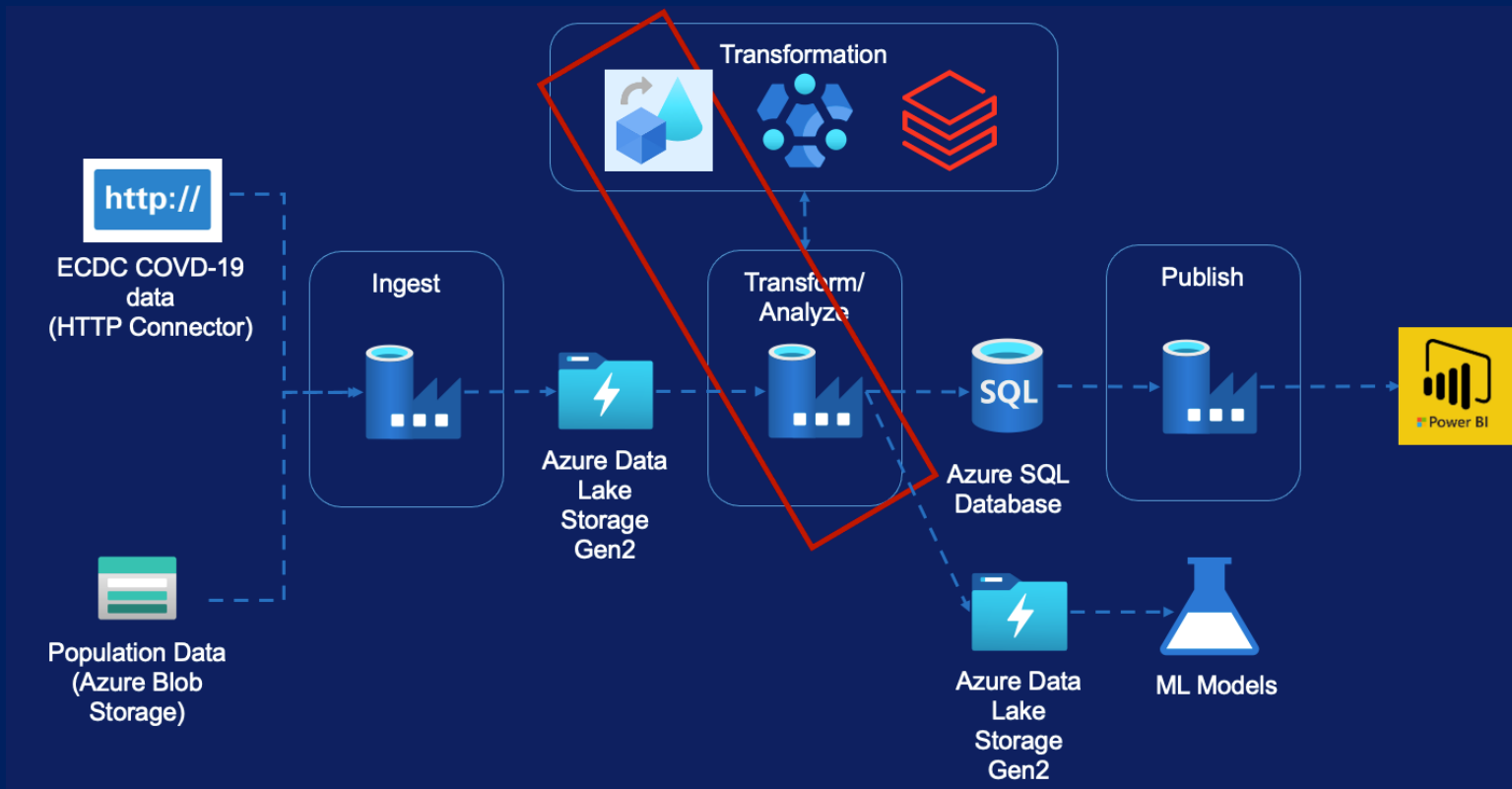


## Transformed File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit
population
cases_count ✓
deaths_count ✓
reported_date(Rename) ✓
source

# Data Flows (2) - Module Overview (Hospital Admissions File)

# Data Flow – Cases & Deaths Data



Requirement

Source Transformation

Select Transformation

Lookup Transformation

Pivot Transformation

Sink Transformation

Conditional Split Transformation

Derived Column Transformation

Aggregate Transformation

Sort Transformation

Join Transformation

Create Pipeline

# Hospital Admissions Data



# Hospital Admissions Data

## Raw File from ECDC

Column Name
country
indicator
date
year_week
value
source
<del>url</del>

## Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

## Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source

# Hospital Admissions Data

## Raw File from ECDC

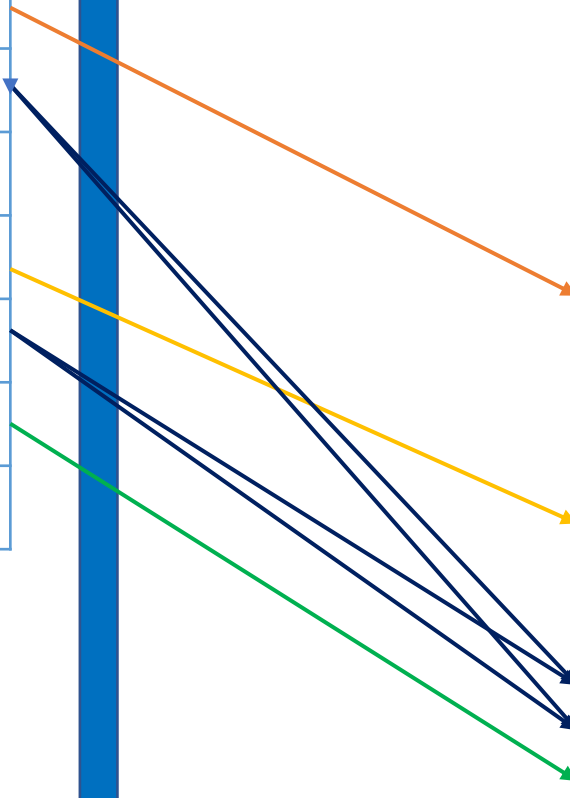
Column Name
country
indicator
date
year_week
value
source
<del>url</del>

## Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

## Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source



# Source Transformation

## Assignment





# Select Transformation Assignment



Remove url



Rename date to reported\_date





Rename year\_week to reported\_year\_week

# Lookup Transformation

## Assignment



-  Lookup country file
-  Select only required fields (i.e. remove additional fields from lookup)

# Pivot Transformation

## Assignment



# Hospital Admissions Data

## Raw File from ECDC

Column Name
country
indicator
date
year_week
value
source
<del>url</del>

## Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

## Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source

# Hospital Admissions Data

## Raw File from ECDC

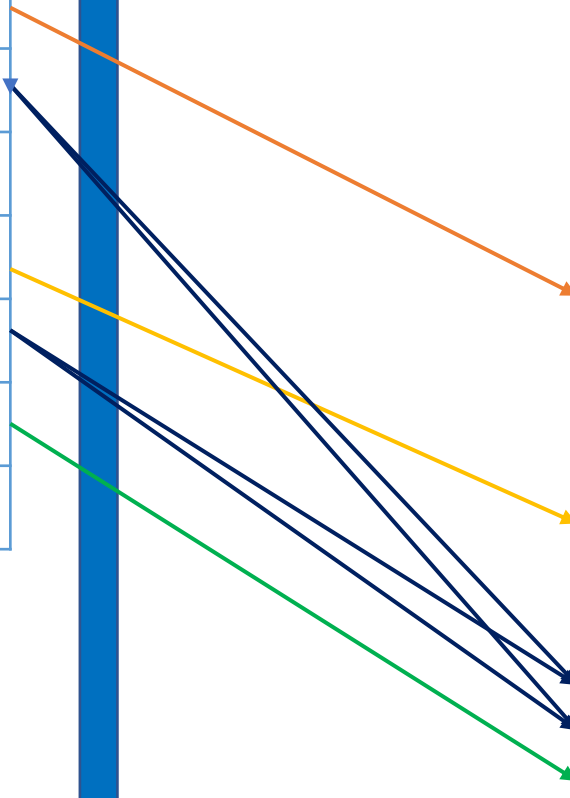
Column Name
country
indicator
date
year_week
value
source
<del>url</del>

## Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

## Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source



# Select & Sink Transformation

## Assignment



# Hospital Admissions Data

## Raw File from ECDC

Column Name
country
indicator
date
year_week
value
source
<del>url</del>

## Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

## Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source

# Hospital Admissions Data

## Raw File from ECDC

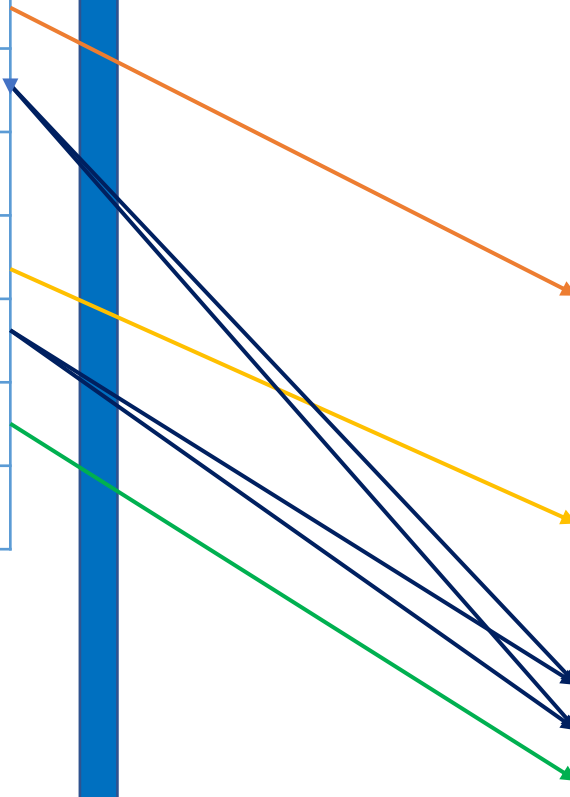
Column Name
country
indicator
date
year_week
value
source
<del>url</del>

## Transformed Daily File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
Population(Lookup)
reported_date
hospital_occupancy_count
icu_occupancy_count
source

## Transformed Weekly File

Column Name
country
country_code_2_digit(Lookup)
country_code_3_digit(Lookup)
population(Lookup)
reported_year_week(transformed)
reported_week_start_date(Lookup)
reported_week_end_date(Lookup)
new_hospital_occupancy_count
new_icu_occupancy_count
Source





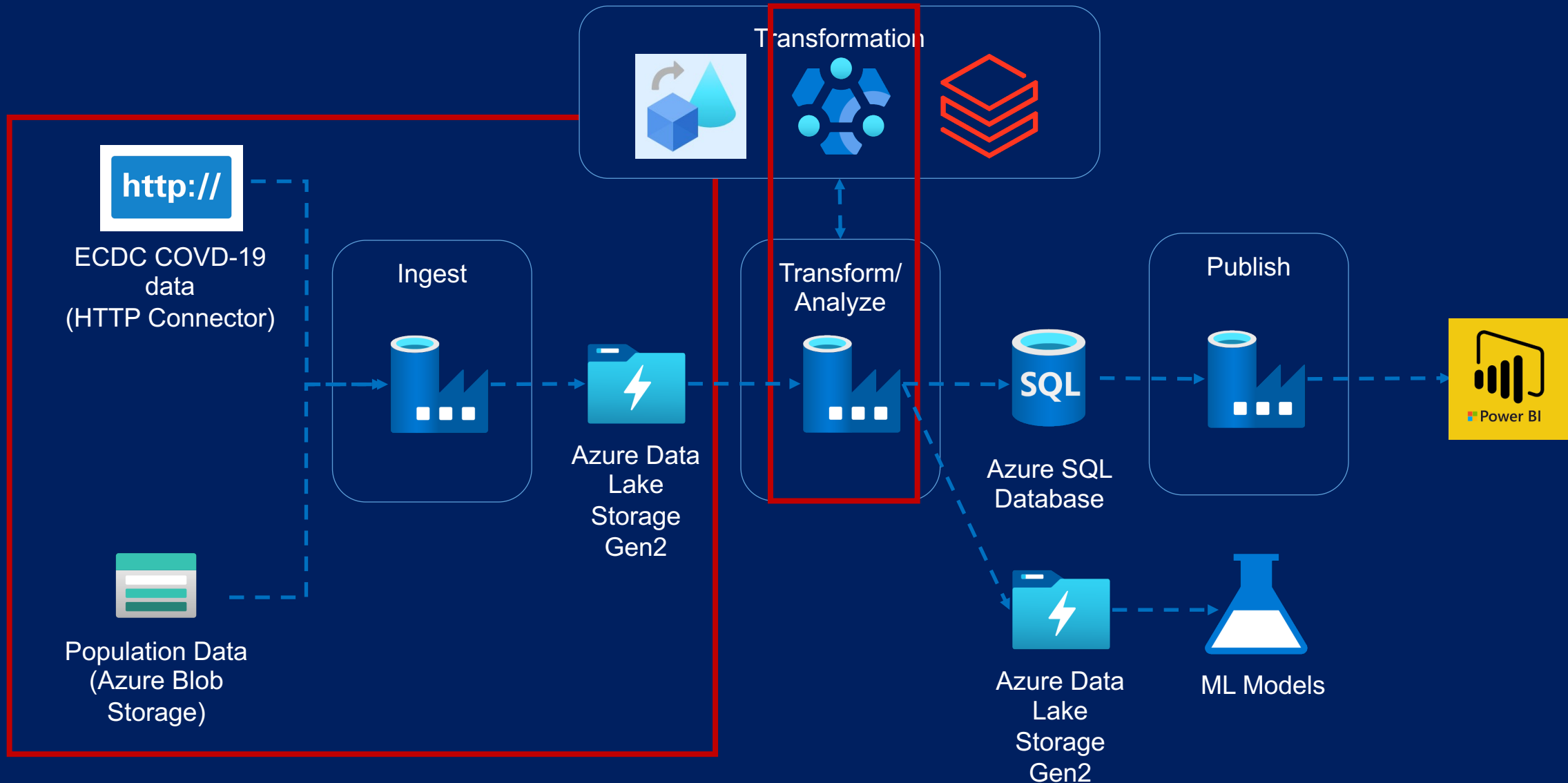
# Data Flow Execution

## Assignment

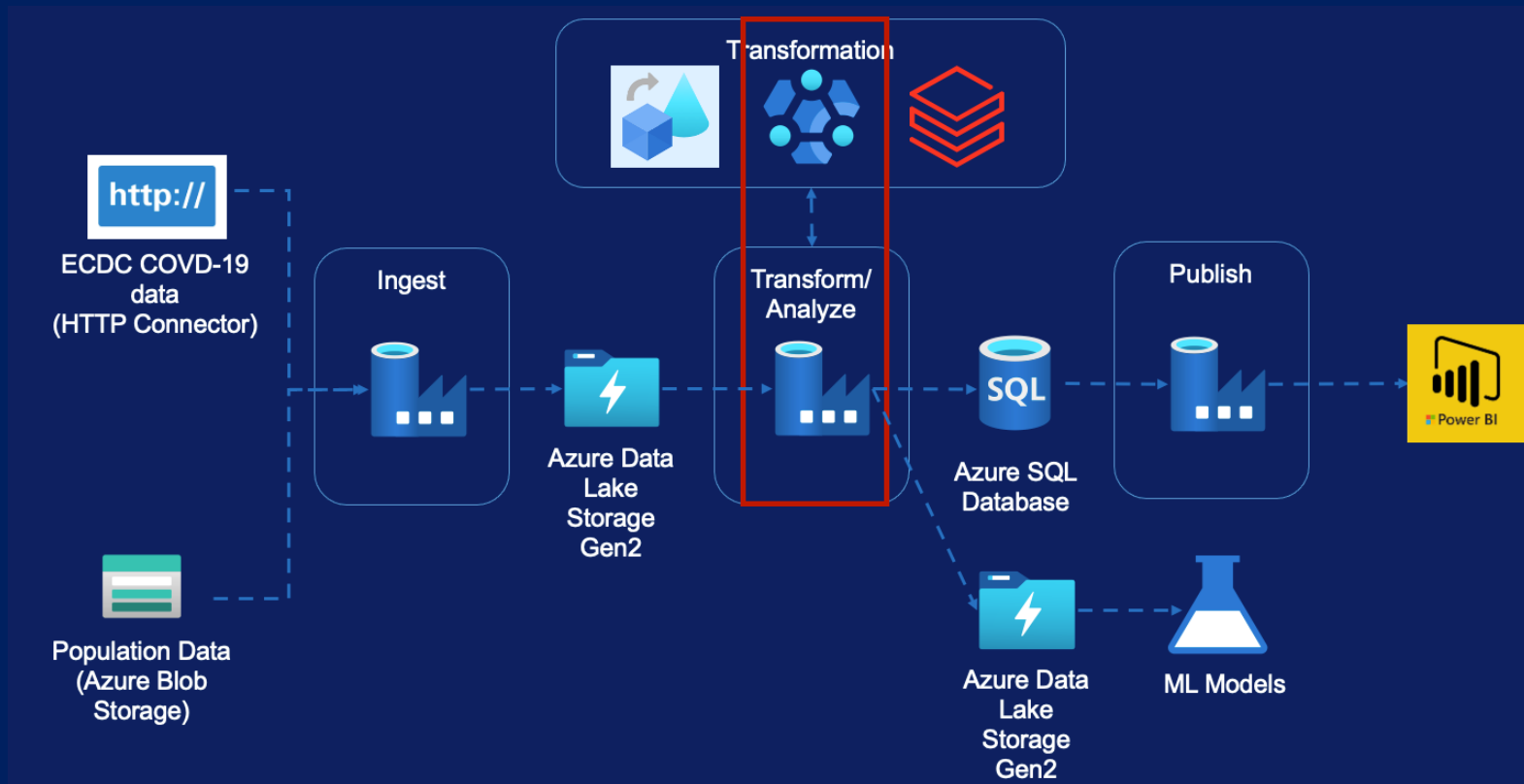


# HDInsight Activity - Module Overview (Testing File)

# HDInsight Activity – Testing File



# HDInsight Activity – Testing File



Creating HDInsight Cluster

HDInsight UI Overview

Transformation Requirement

Hive Script Walk-through

Creating Pipeline

Delete HDInsight Cluster

# Creating HDInsight Cluster



# Testing Data



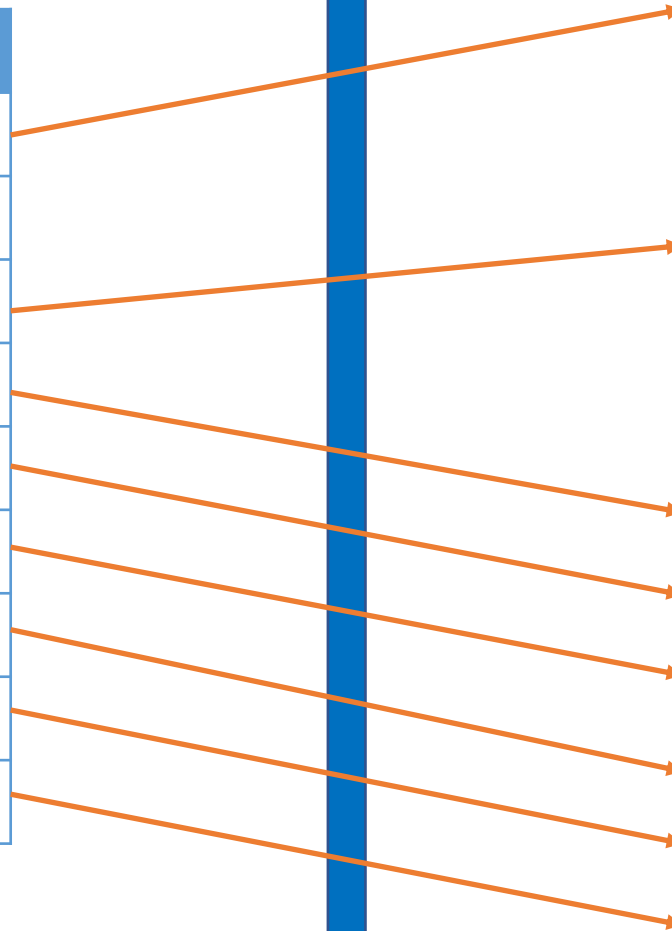
# Testing Data

## Raw File from ECDC

Column Name
country
country_code (Remove)
Year_week
new_cases
test_done
population
testing_rate
positivity_rate
testing_data_source

## Transformed File

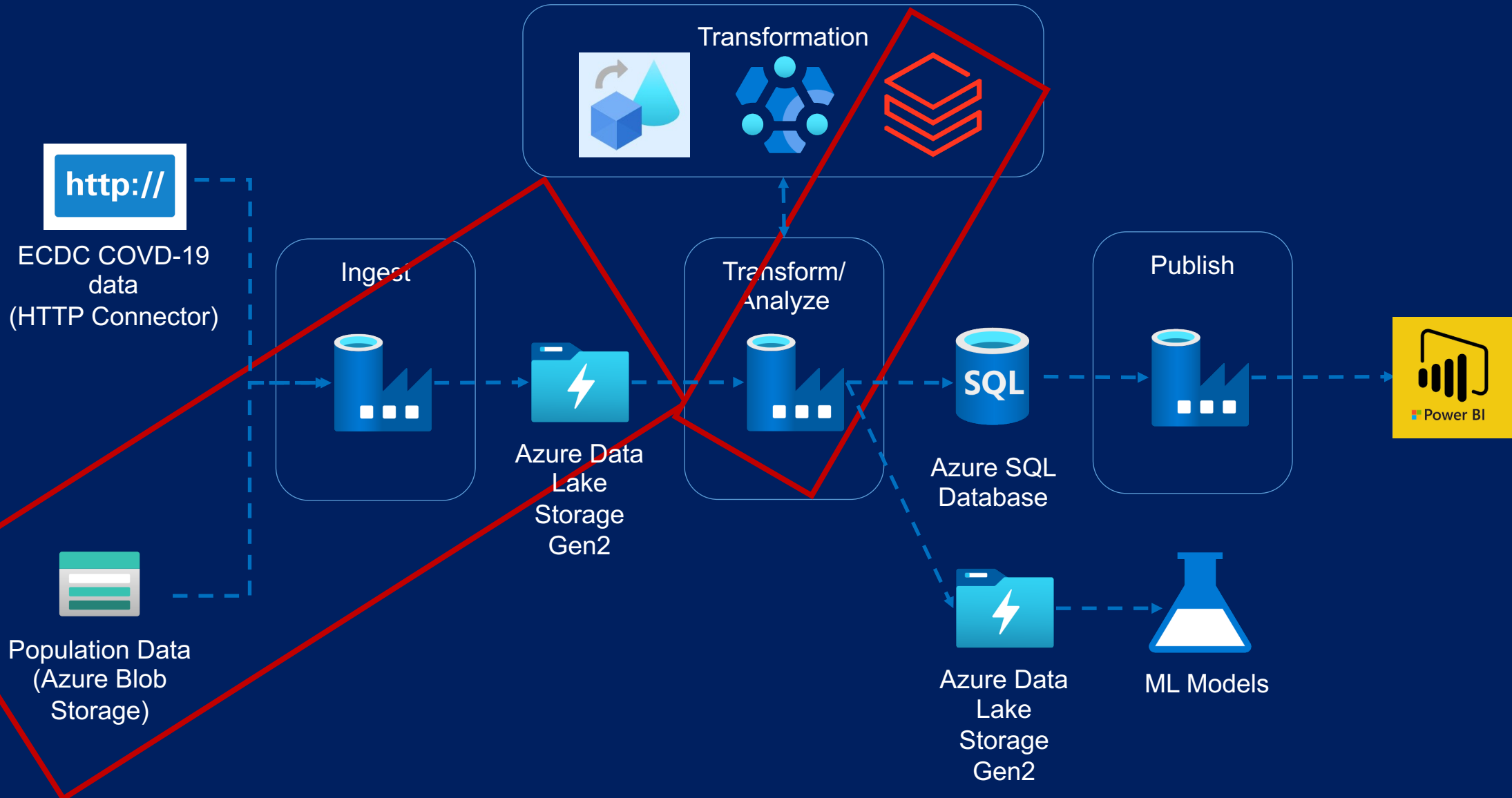
Column Name
country
country_code_2_digit (lookup)
country_code_3_digit(lookup)
reported_year_week
reported_week_start_date(lookup)
reported_week_end_date(lookup)
new_cases
test_done
population
testing_rate
positivity_rate
testing_data_source



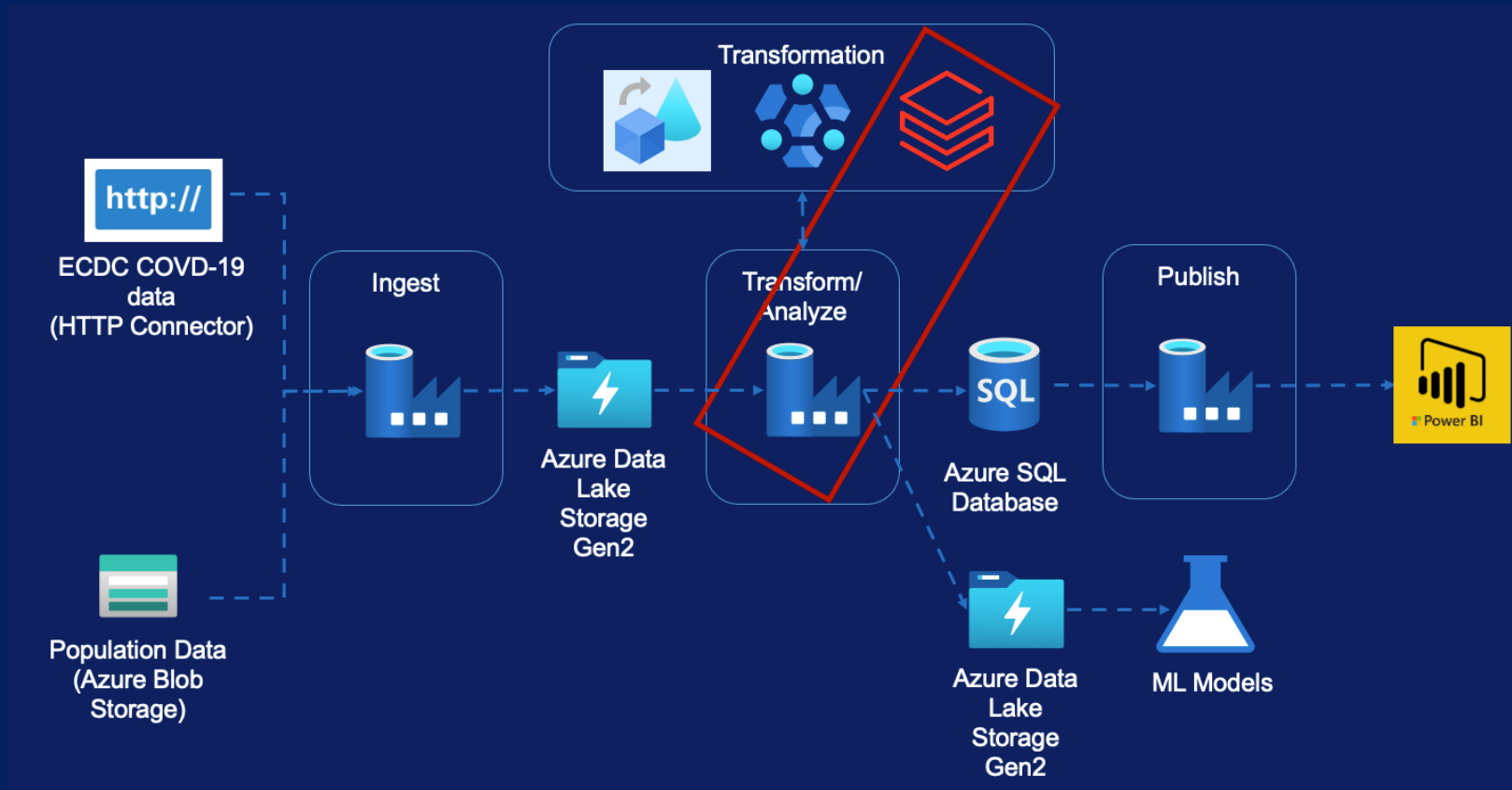
# Databricks Activity - Module Overview (Population File)



# Databricks Activity – Population File



# Databricks Activity – Population File



Create Databricks Service

Create Databricks Cluster

Mount Storage Accounts

Transformation Requirements

Creating Pipeline

# Databricks Environment Set-up



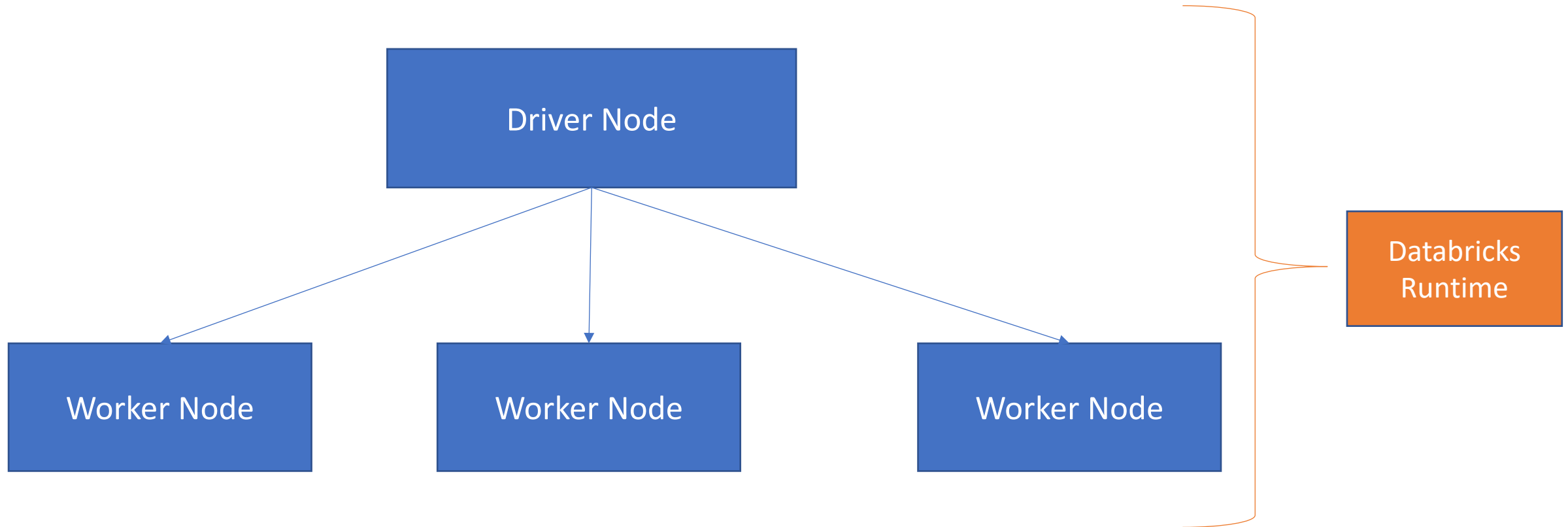
# Creating Databricks Service



# Creating Databricks Cluster



# What is a cluster?



# Cluster Types

All Purpose/ Interactive  
Clusters

Job Clusters

# Mounting Data Lake Storage





# Mounting Data Lake Storage

- Create Azure Service Principal
- Grant access for data lake to Azure Service Principal
- Create the mount in databricks using Service Principal

# Transform Population By Age Data



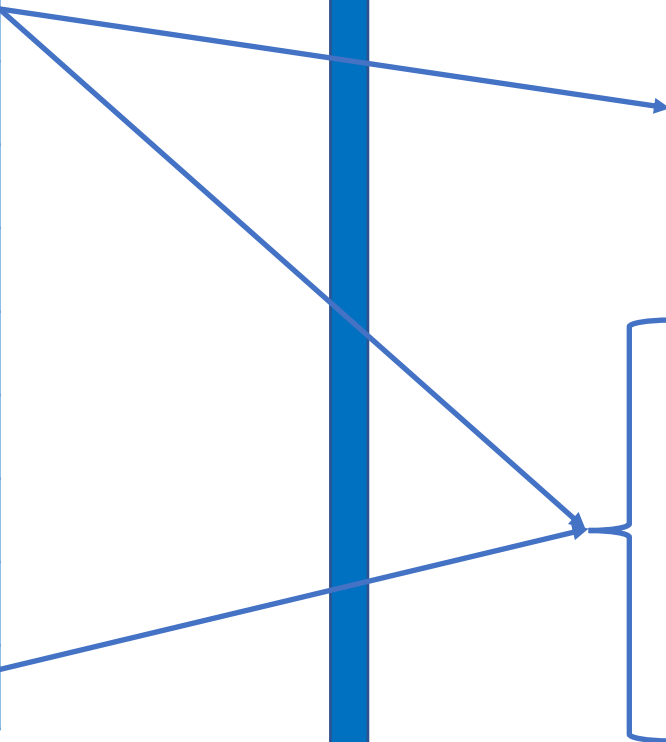
# Transform Population By Age Data

Raw File

Column Name
indic_de,geo\time
2008
2009
2010
2011
...
....
2018
2019

Transformed File

Column Name
Country (Lookup)
country_code_2_digit(Substr)
country_code_3_digit(Lookup)
population(Lookup)
age_group_0_14
age_group_25_49
age_group_50_64
age_group_65_79
age_group_80_max



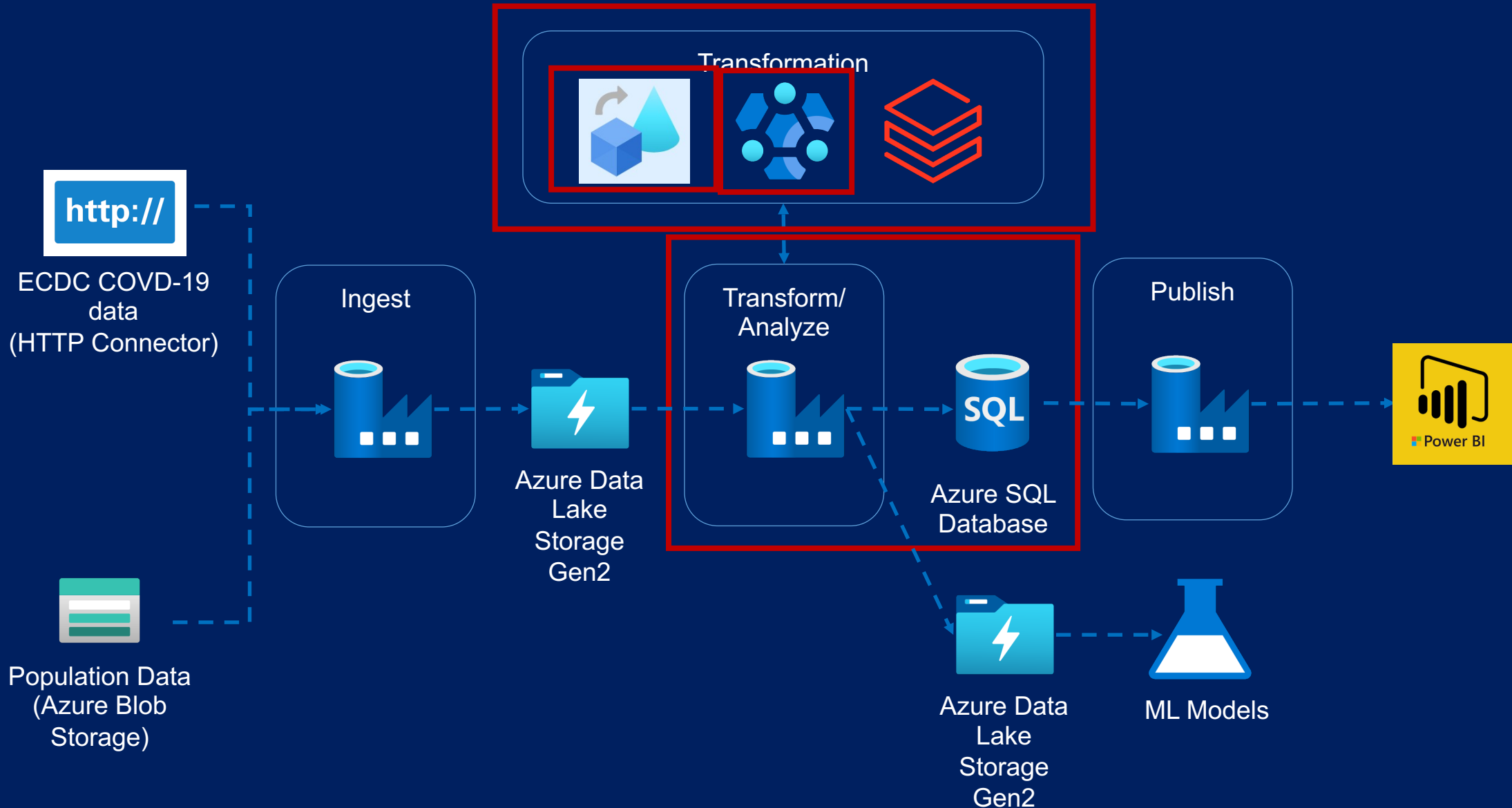
# Transform Population By Age Data

*Data Factory Pipeline*

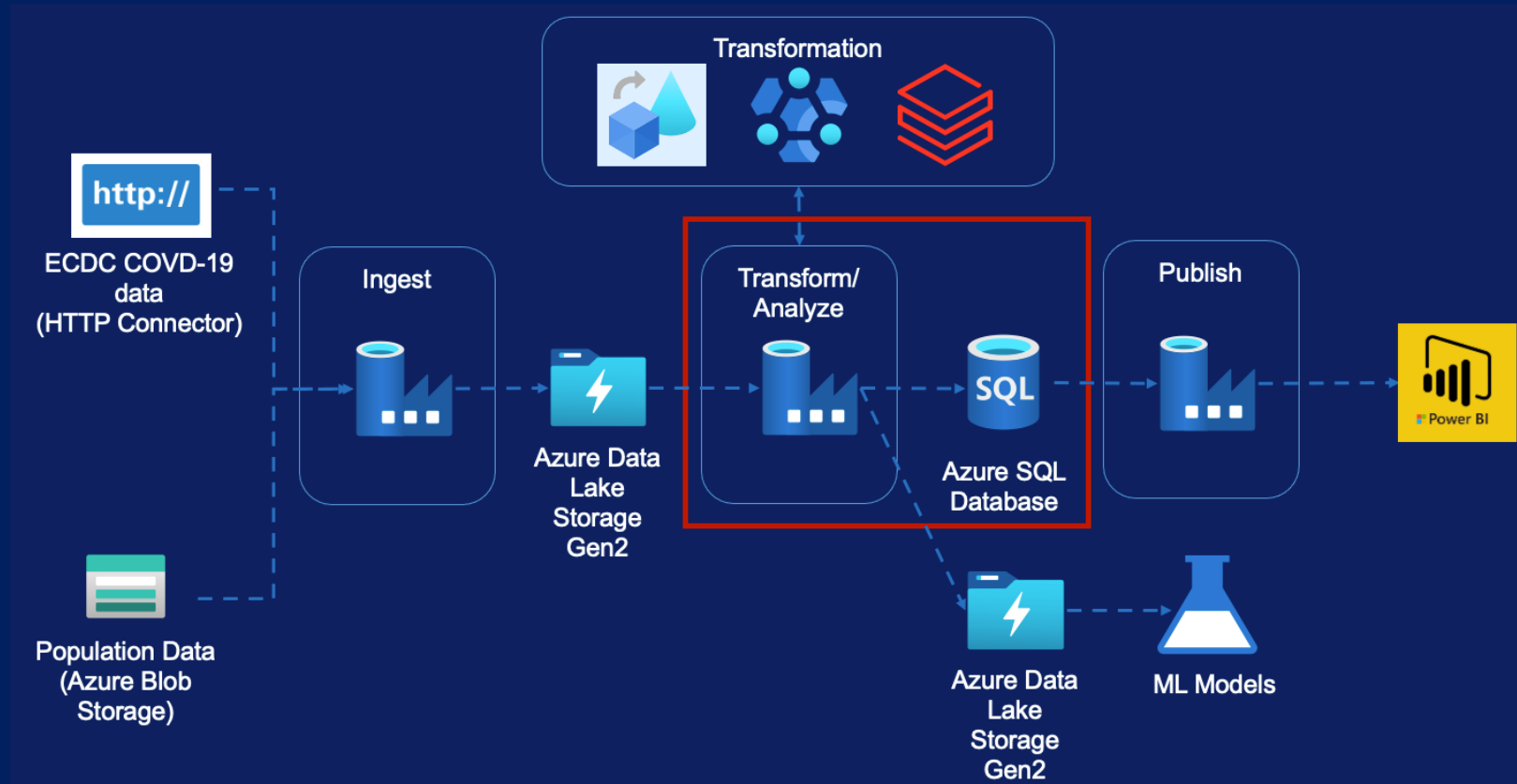


Copy Data to Azure SQL

# Copy Data to SQL



# Copy Data to SQL



- Copy Cases & Deaths
- Copy Hospital Admissions
- Copy Testing

# Copy Activity – Data Lake to SQL

Cases and Deaths Data





# Copy Activity – Data Lake to SQL

Hospital Admissions Daily Data



## Assignment

# Copy Activity – Data Lake to SQL

Testing Data



# Data Orchestration



# Data Orchestration Requirements

- Pipeline executions are full automated
- Pipelines run at regular intervals or on an event occurring
- Activities only run once the upstream dependency has been satisfied
- Easier to monitor for execution progress and issues

# Data Factory Capability

- Dependency between activities inside a pipeline
- Dependency between pipelines within a parent pipeline
- Dependency between triggers [Only tumbling window triggers]
- Custom-made Solution

# Data Orchestration

Option 1 – Parent Pipeline



# Data Orchestration

Option 2 – Trigger Dependency



# Azure Data Factory - Monitoring



# Azure Data Factory - Monitoring



- What to Monitor
- Data Factory Monitoring
- Creating Alerts
- Recovery From Failure
- Reporting on Metrics
- Azure Monitor Introduction
- Log Analytics
- Azure Data Factory Analytics

# Monitoring

# What do we want to monitor

- Azure Data Factory Resource

- Integration runtime

- Trigger runs

- Pipeline runs

- Activity runs

# Data Factory Monitor

- Ability to monitor status of pipeline/ triggers
- Can be used to re-run failed pipelines/ triggers
- Ability to send alerts from base level metrics
- Provides base level metrics and logs
- Pipeline runs are stored only for 45 days

# Azure Monitor

- Ability to route the diagnostic data to other storage solutions
- Provides richer diagnostic data
- Ability to write complex queries and custom reporting
- Ability to report across multiple data factories

# Data Factory Monitor



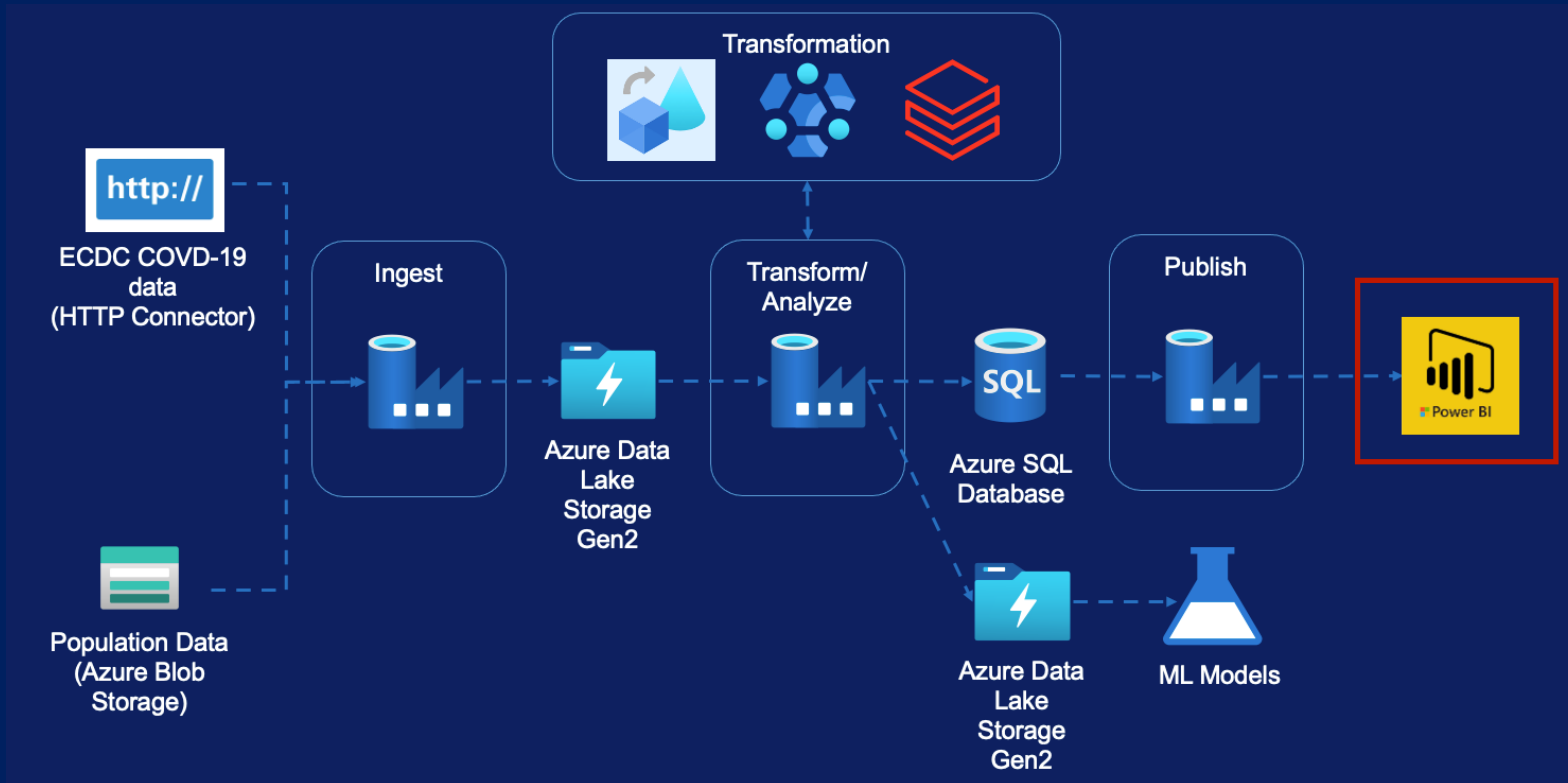
# Azure Monitor



# Reporting via Power BI



# Reporting via Power BI



Introduction to Power BI Desktop

Review the Covid-19 pre-built Report

# Power BI Desktop Overview



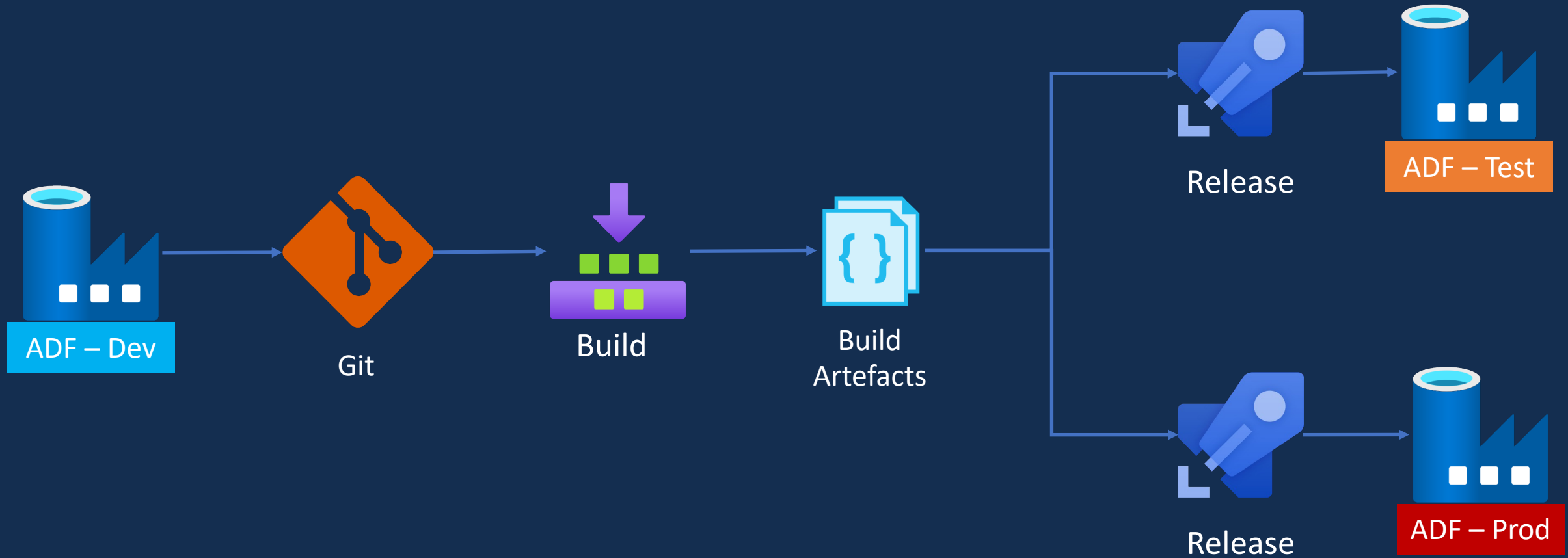
# Continuous Integration / Continuous Delivery (CI / CD)

## *Module Overview*

# Continuous Integration / Continuous Delivery

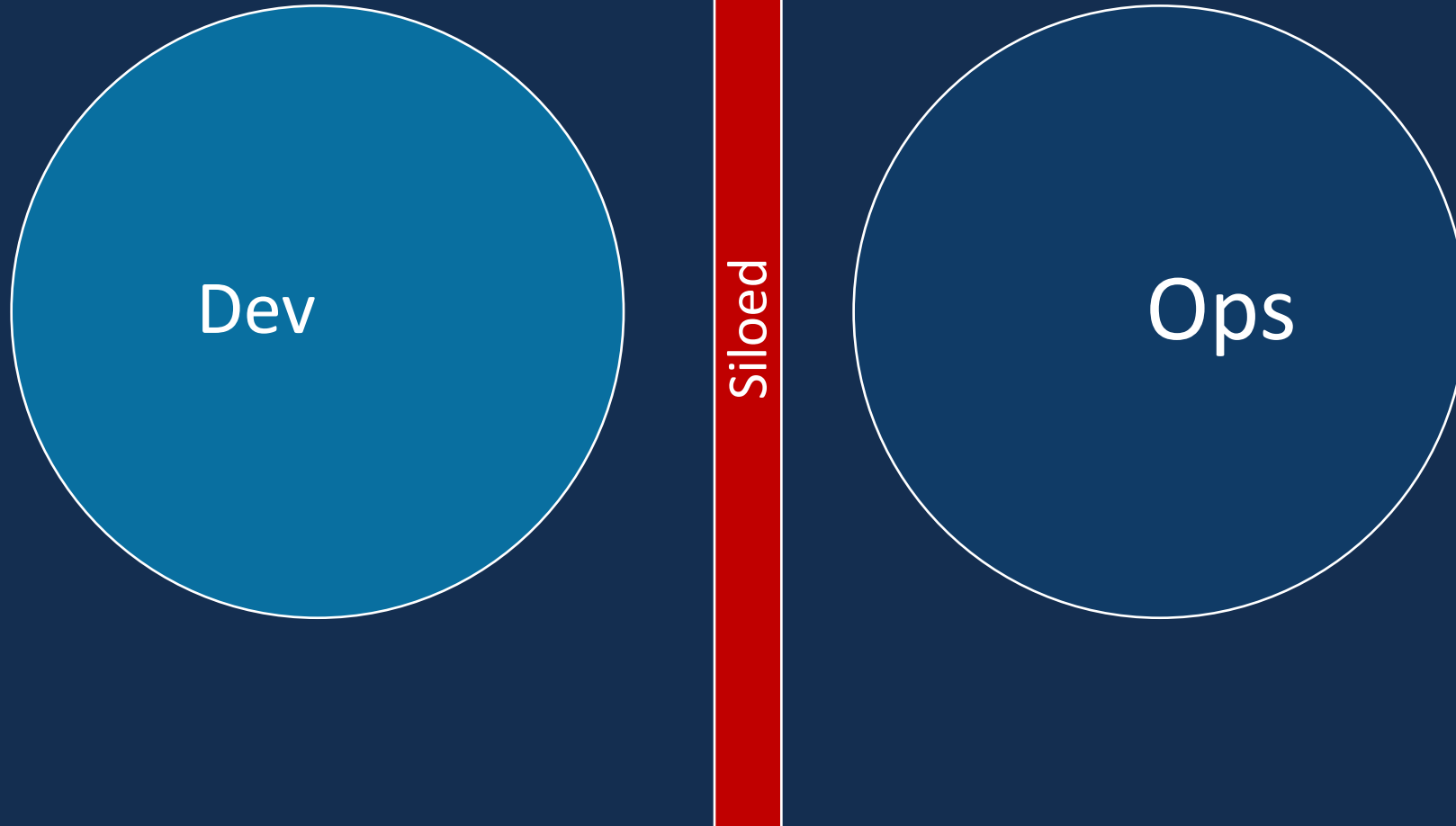


# Continuous Integration / Continuous Delivery

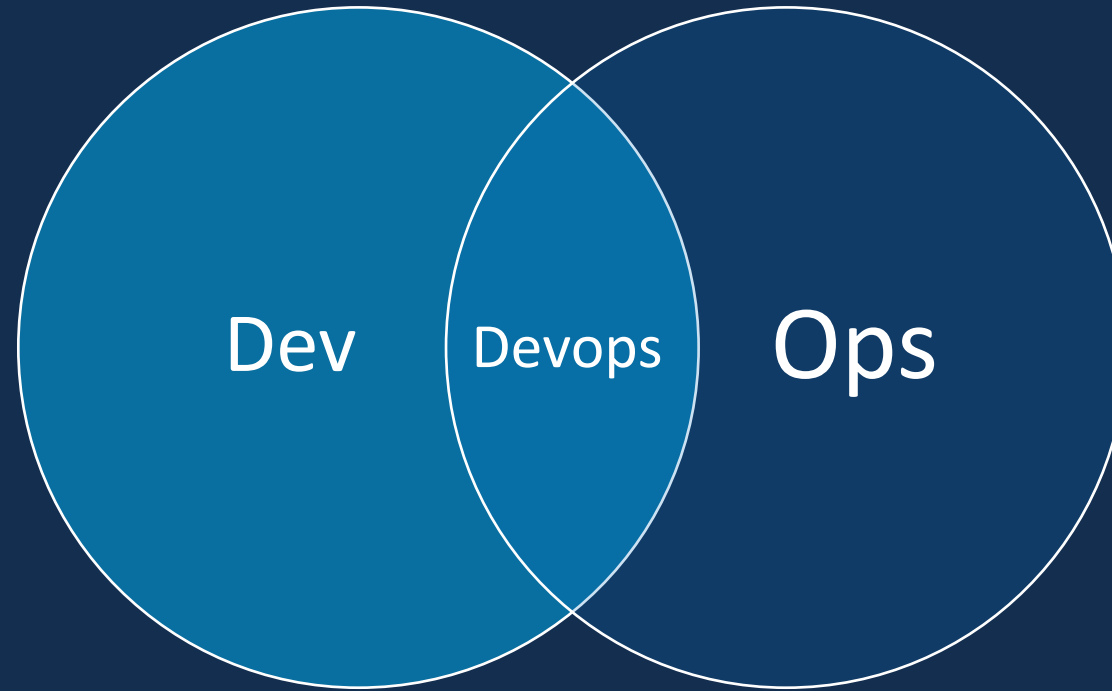


# Continuous Integration / Continuous Delivery (CI / CD)

# DevOps - Introduction

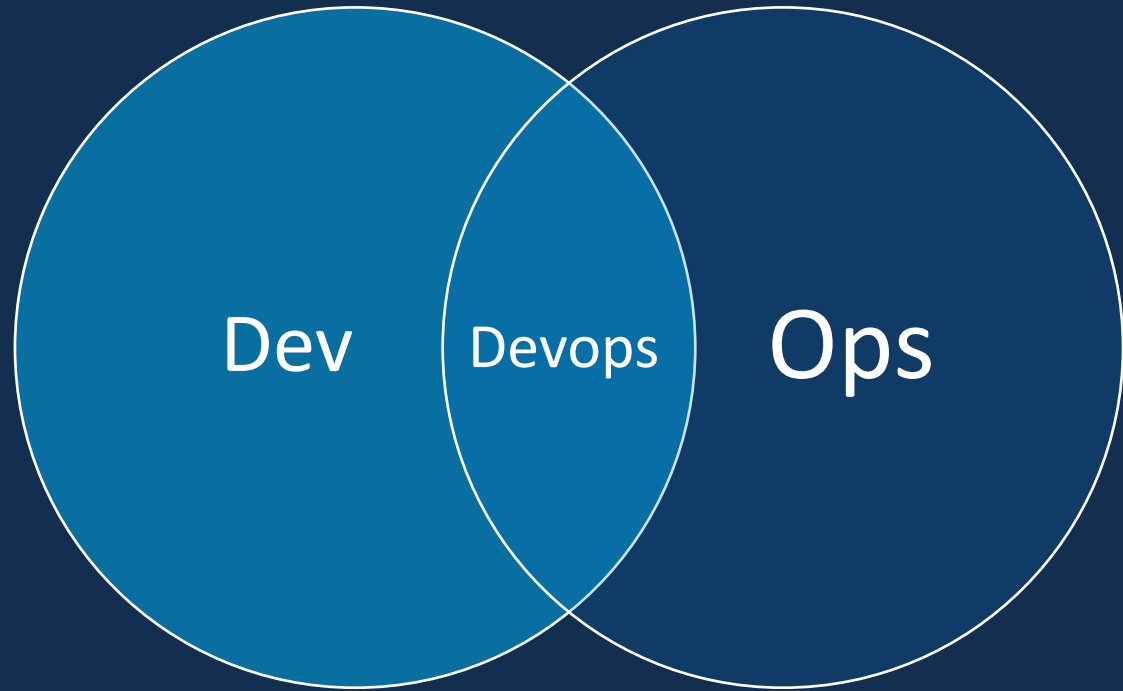


# DevOps - Introduction





# DevOps - Characteristics



Collaboration, trust and transparency

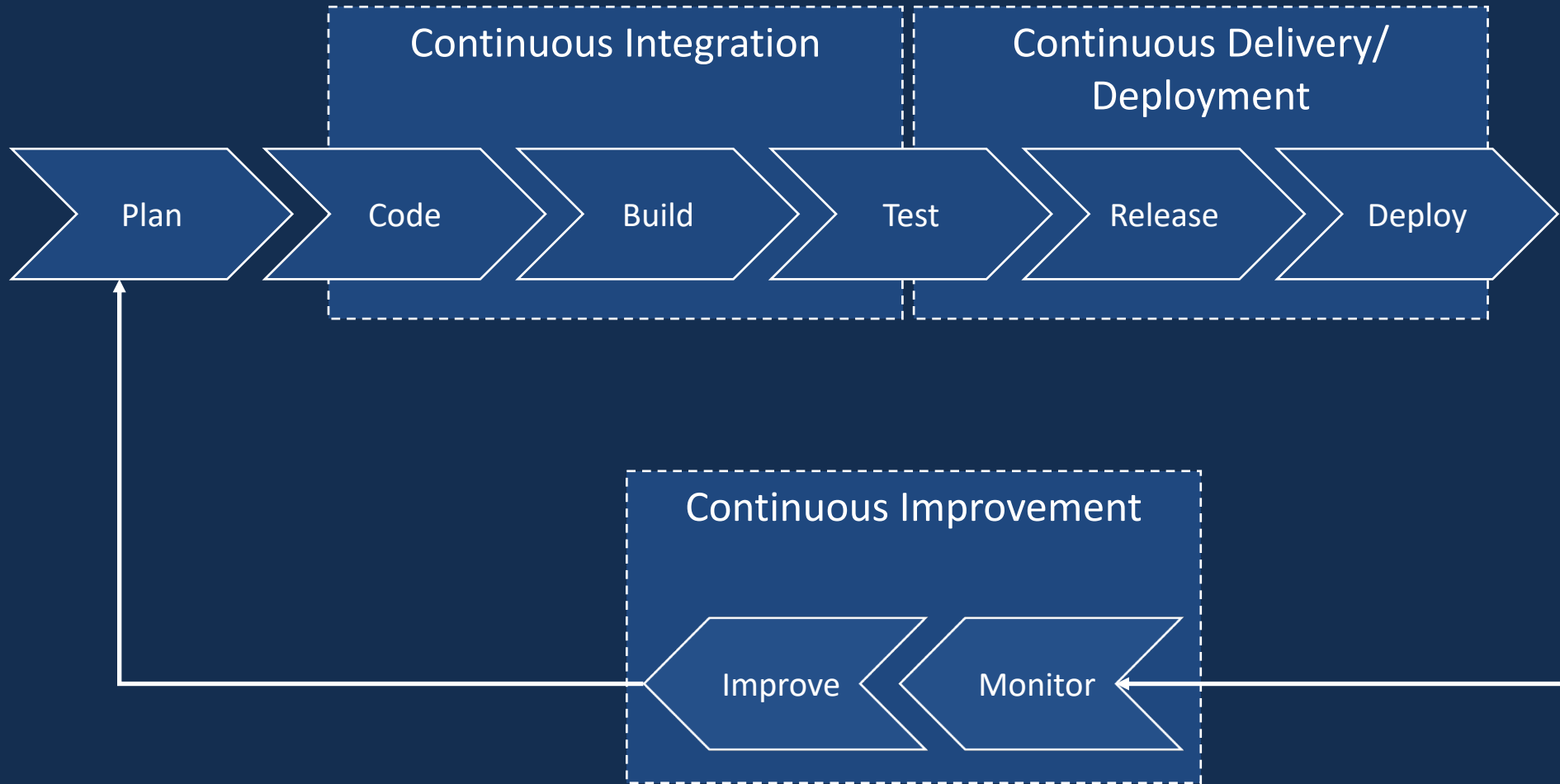
Agile Development Approach

Continuous Integration/ Delivery

Automation

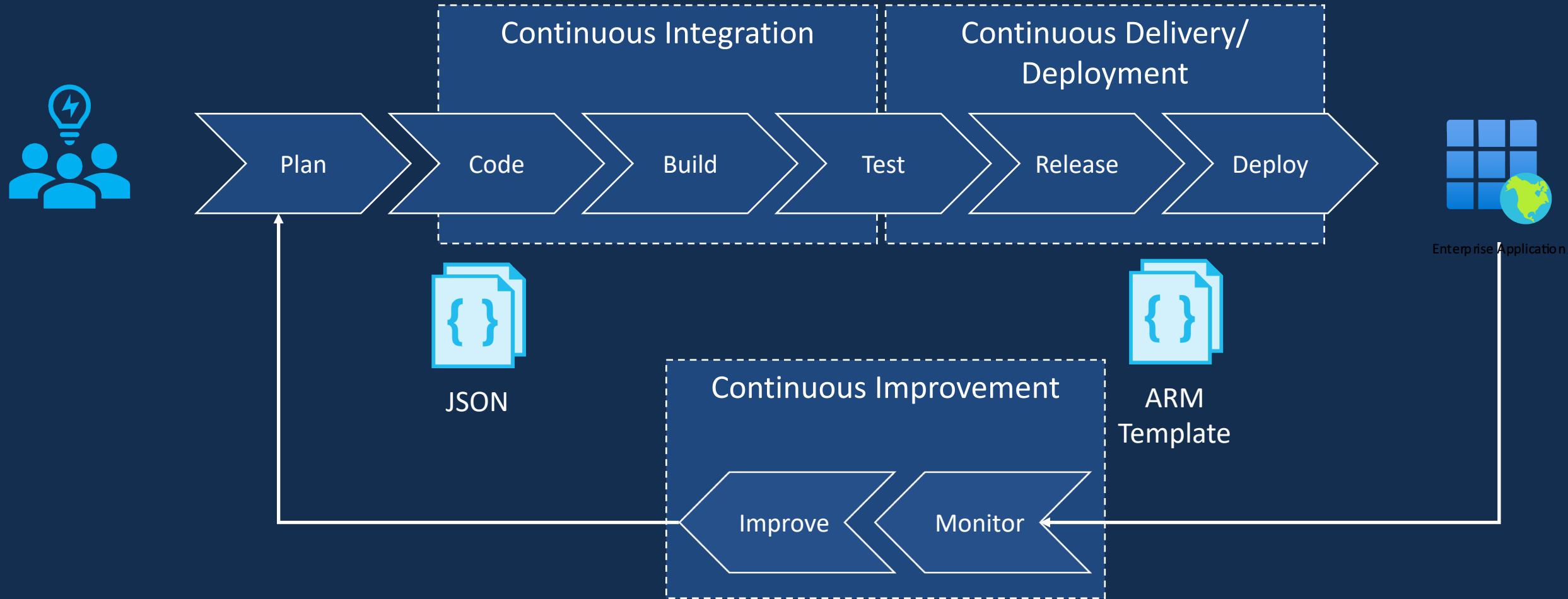
Continuous Improvement

# Continuous Integration / Continuous Delivery

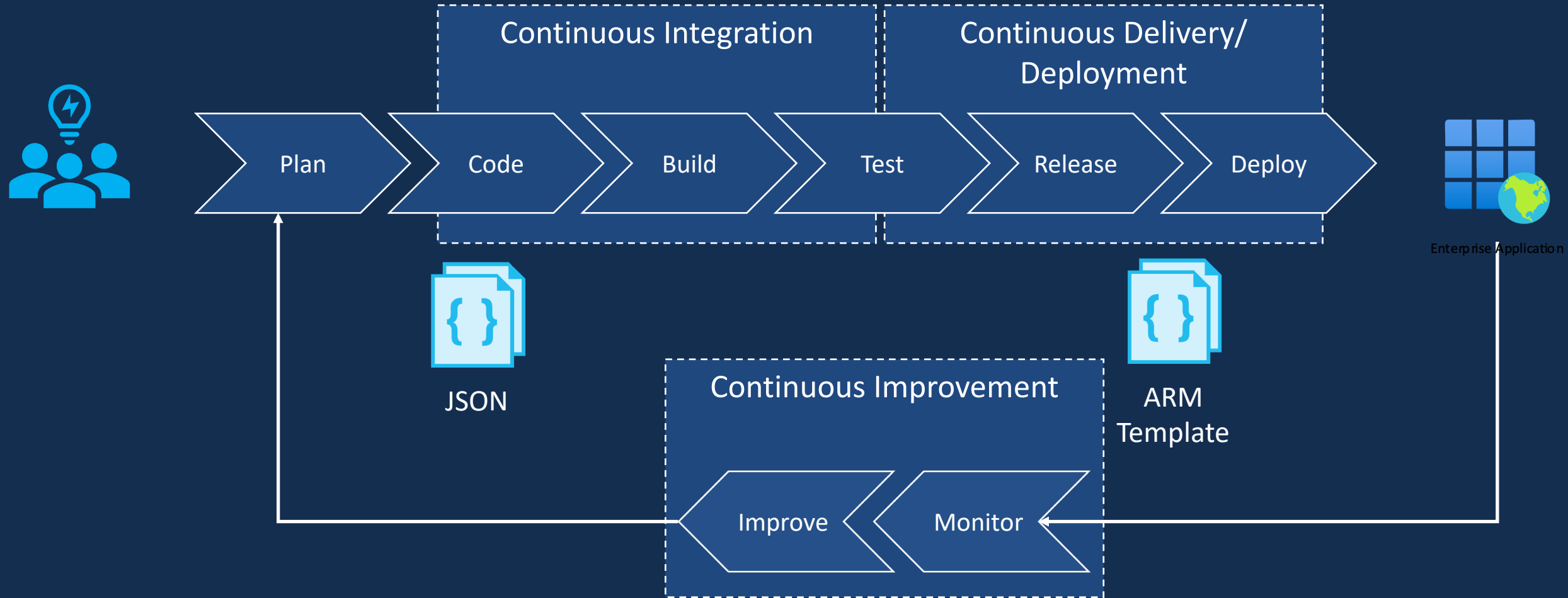


Enterprise Application

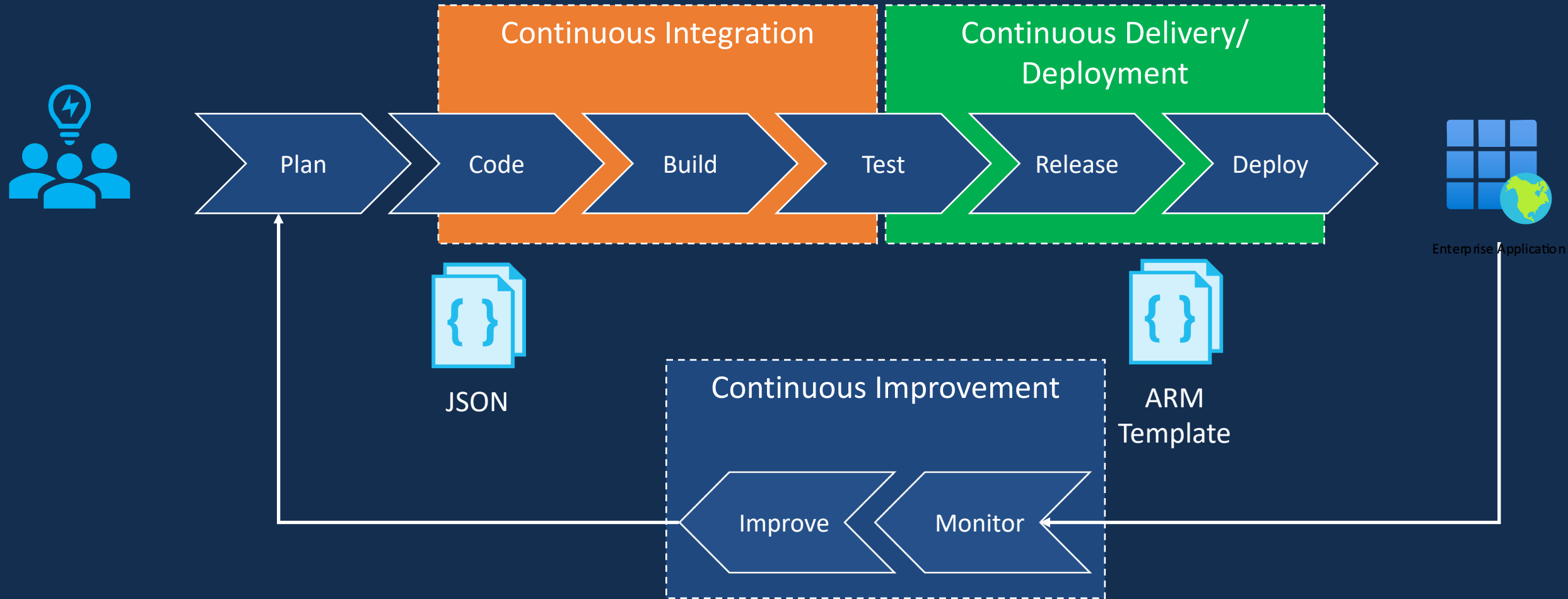
# Continuous Integration / Continuous Delivery - ADF



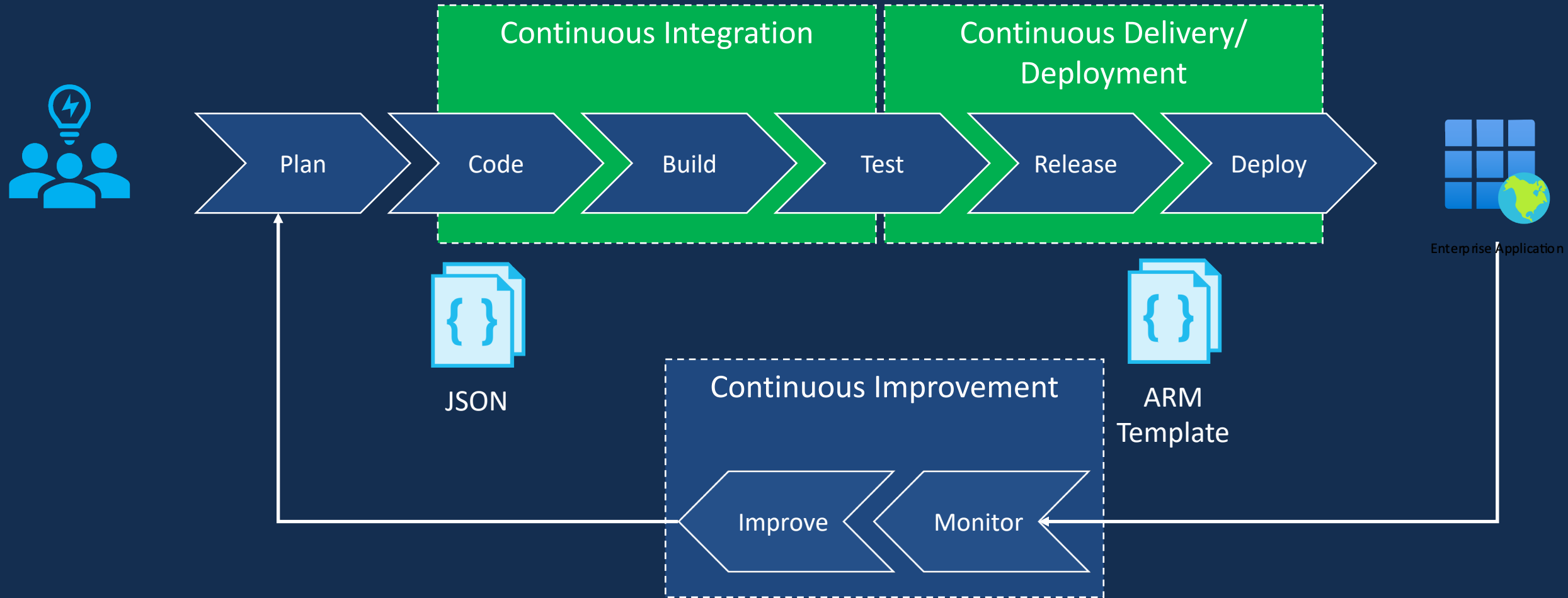
# Continuous Integration / Continuous Delivery - ADF



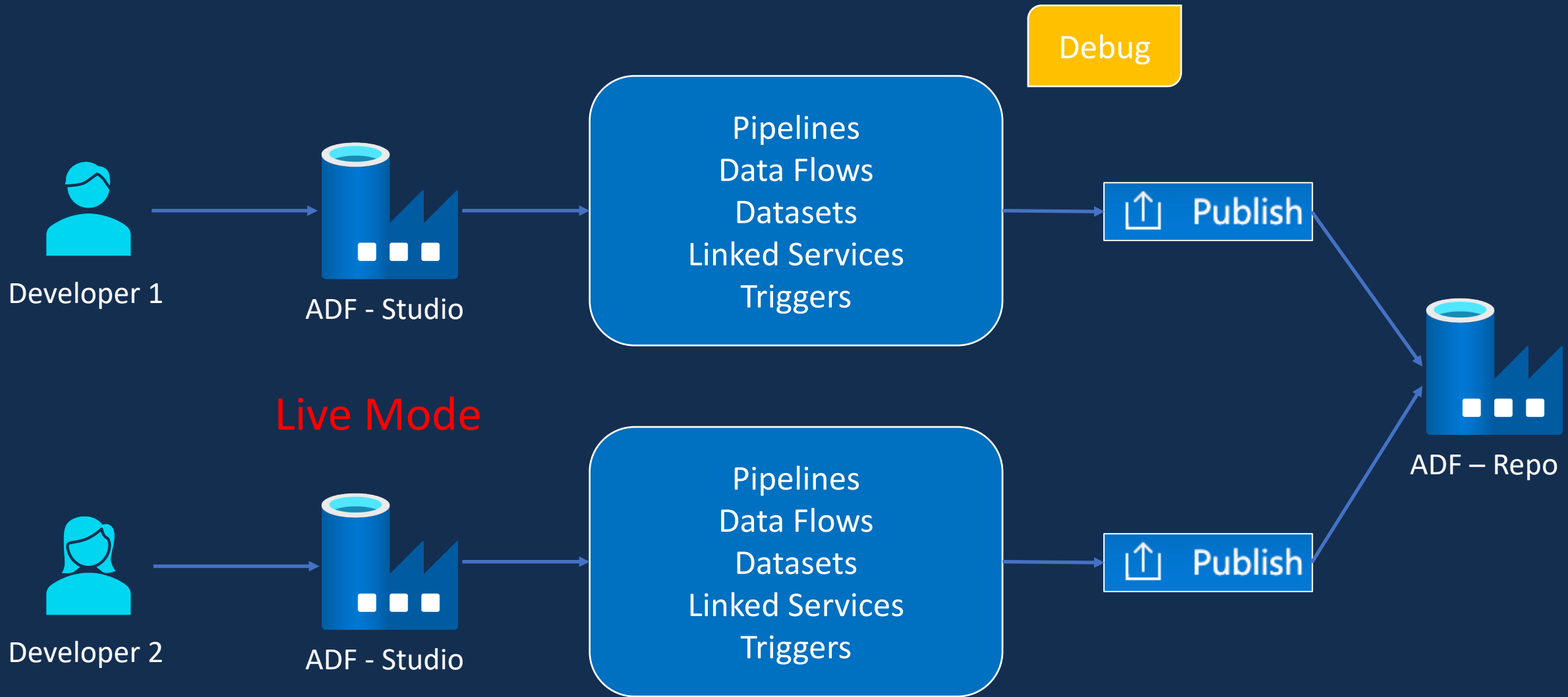
# Option 1 – Manual Integration / Automated Delivery



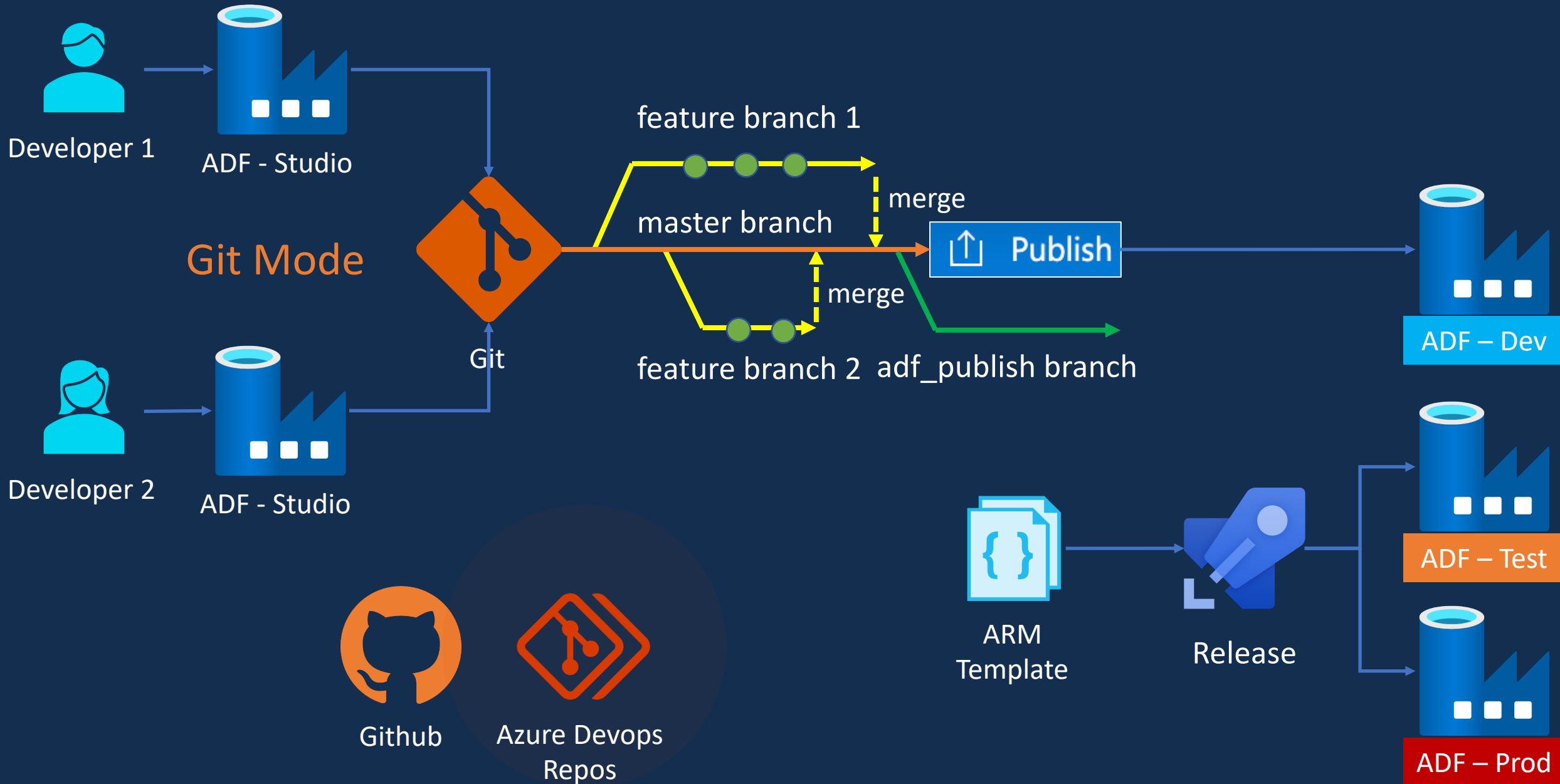
# Option 2 – Fully Automated Solution



# CI/CD Option 1 – Using ADF Publish

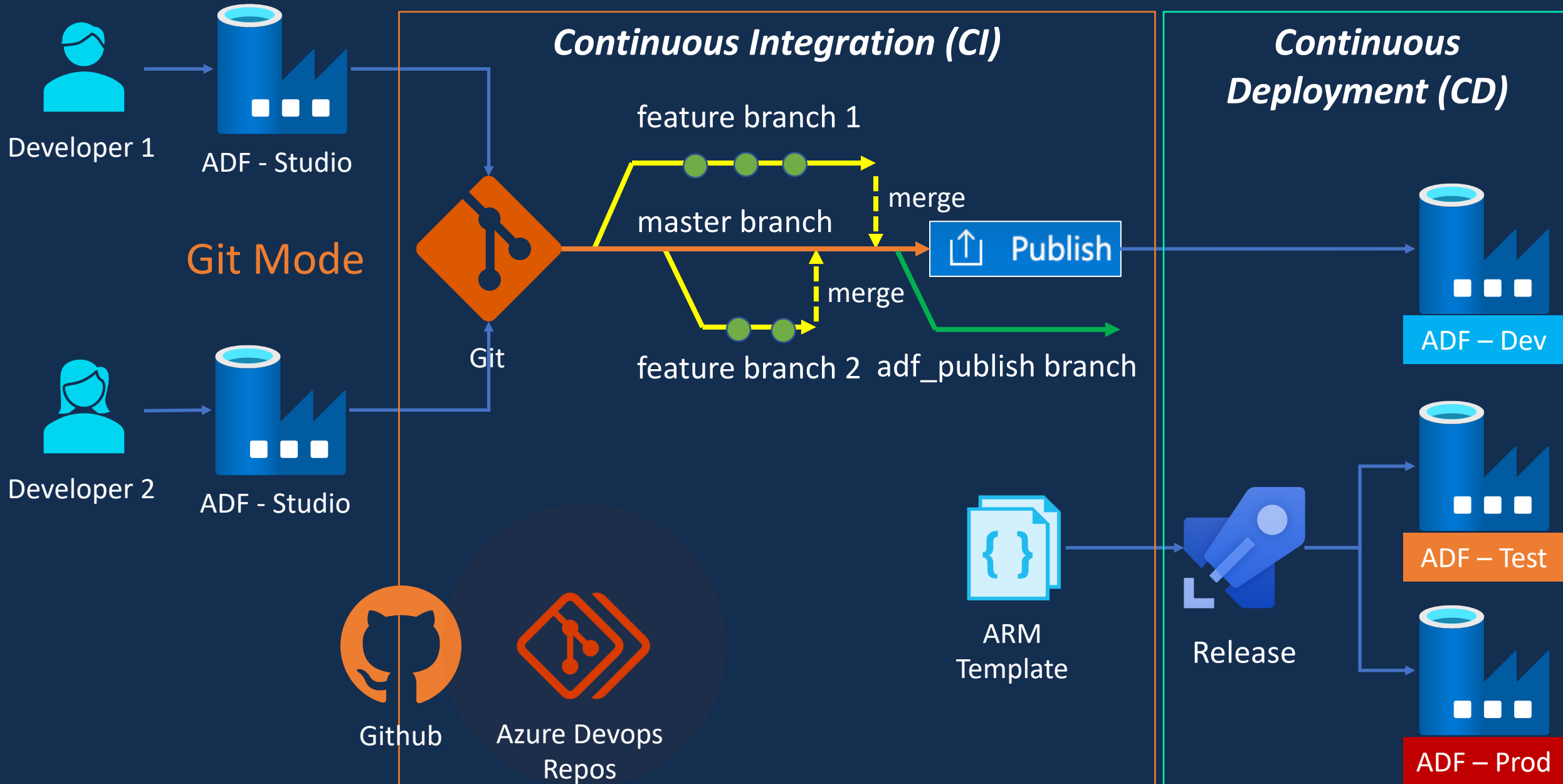


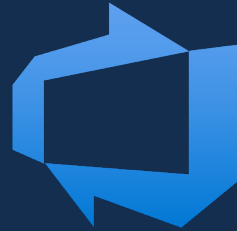
# CI/CD Option 1 – Using ADF Publish





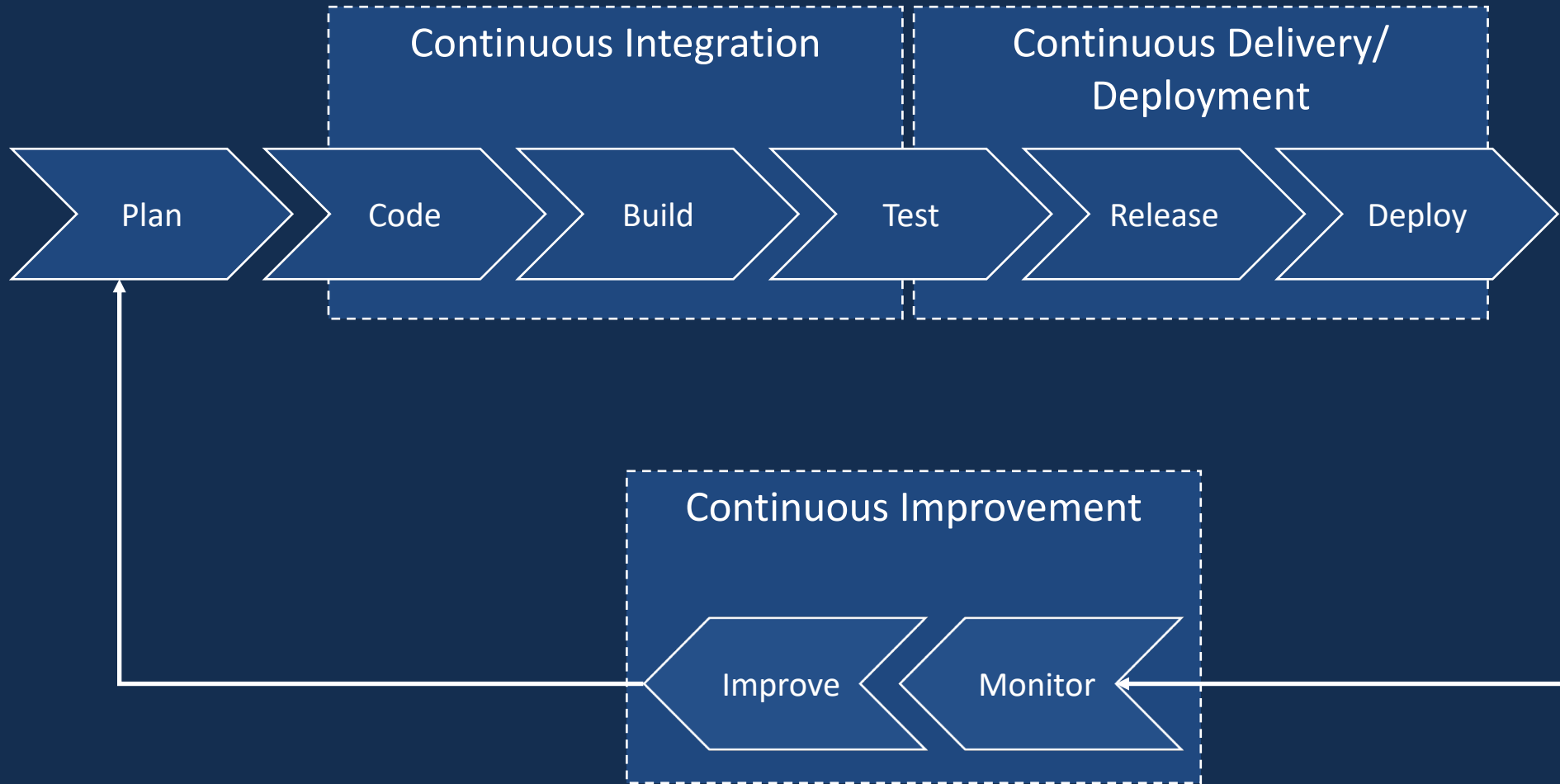
# CI/CD Option 1 – Using ADF Publish





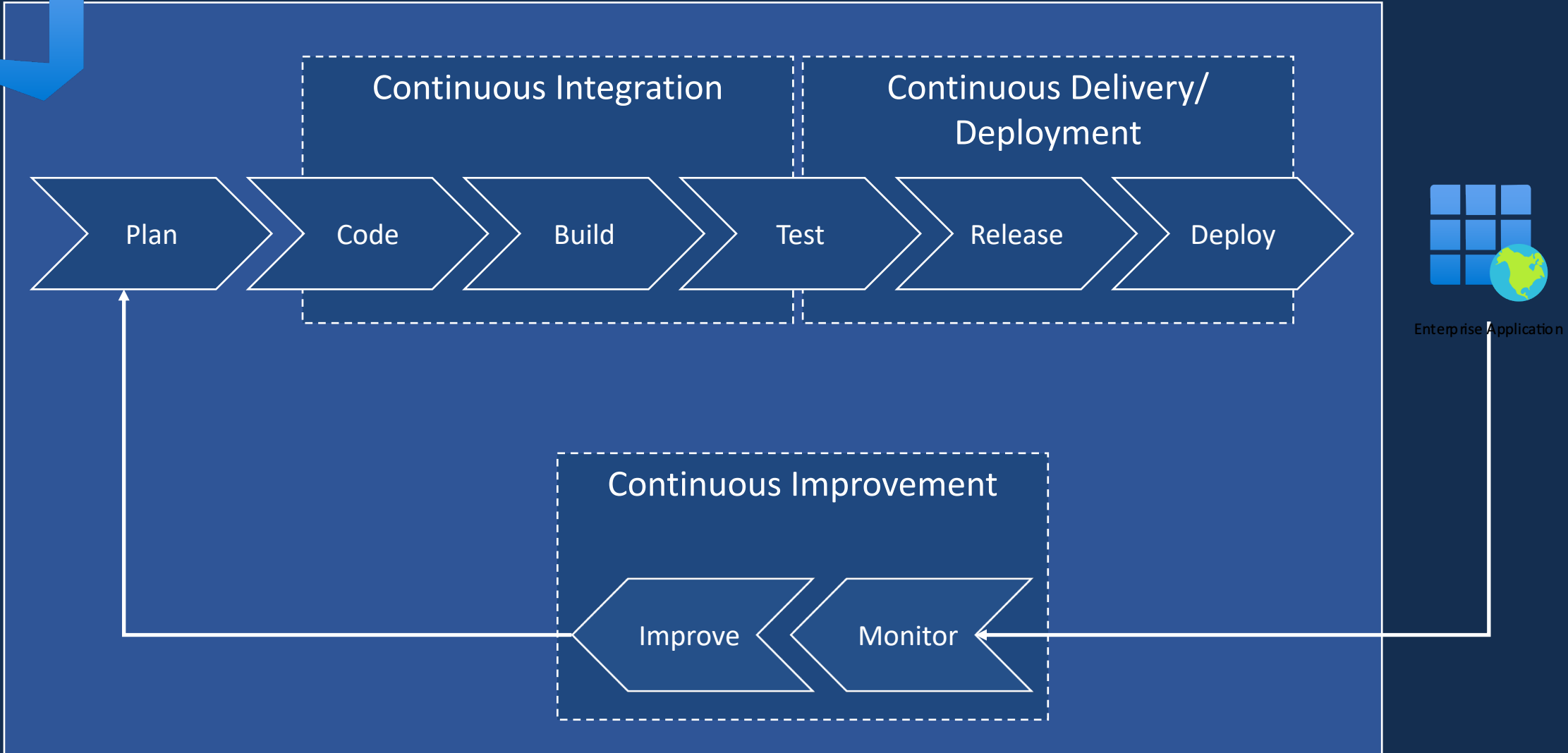
Azure DevOps

# DevOps



Enterprise Application

# Azure DevOps



# Azure DevOps



Boards



Repos



Pipelines

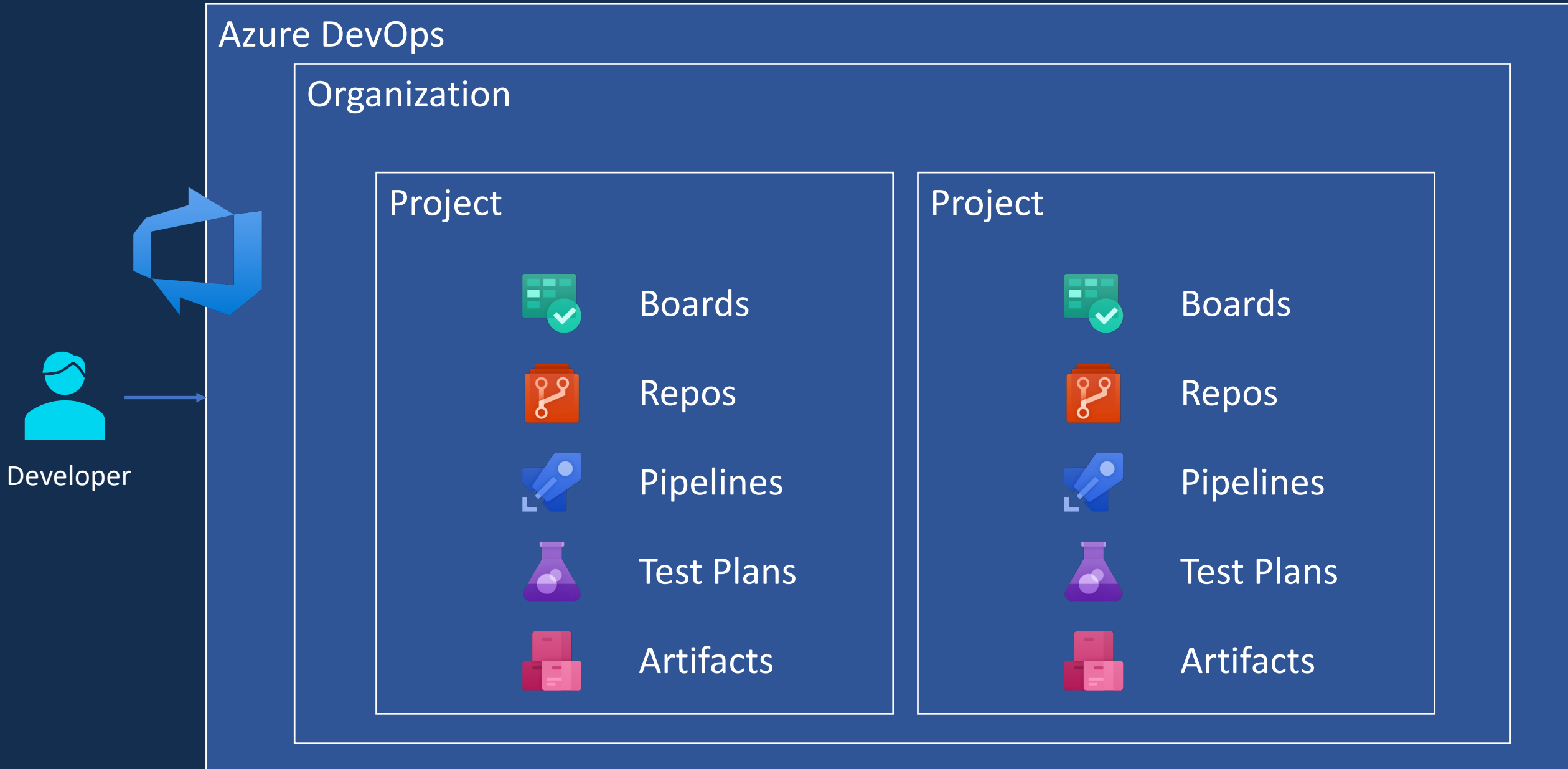


Test Plans



Artifacts

# Azure DevOps



# Azure DevOps Environment Set-up

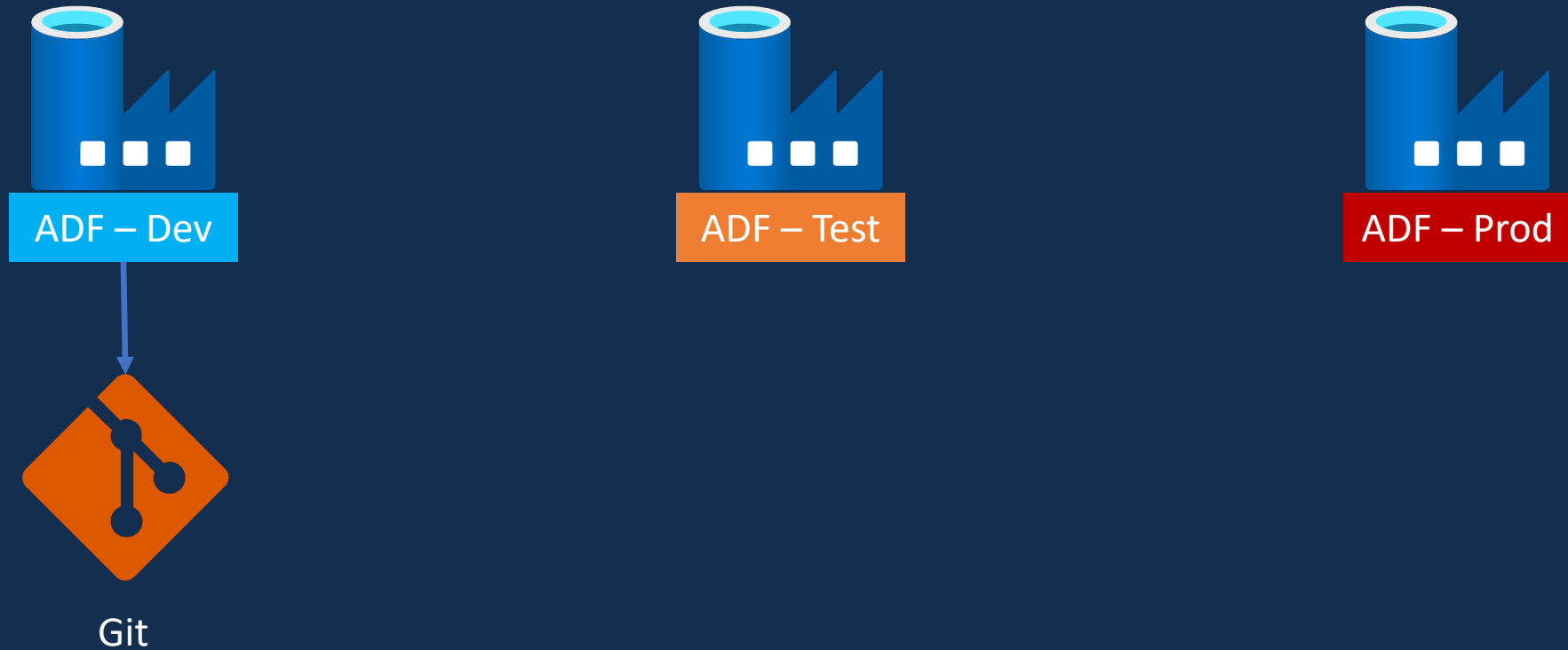


# Azure Data Factory Set-up

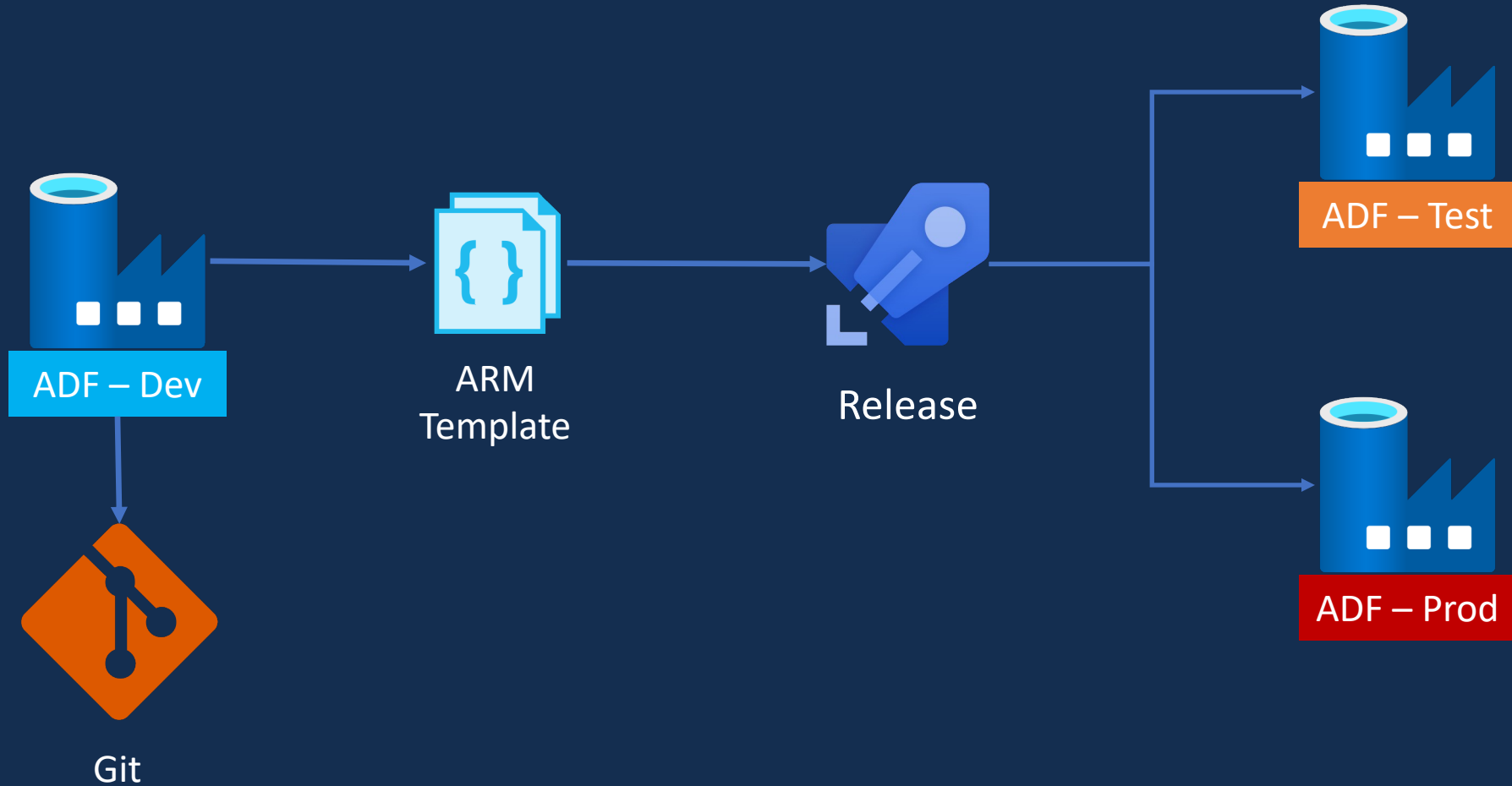




# Azure Data Factory Set-up



# Azure Data Factory Set-up



# Azure Data Factory Set-up

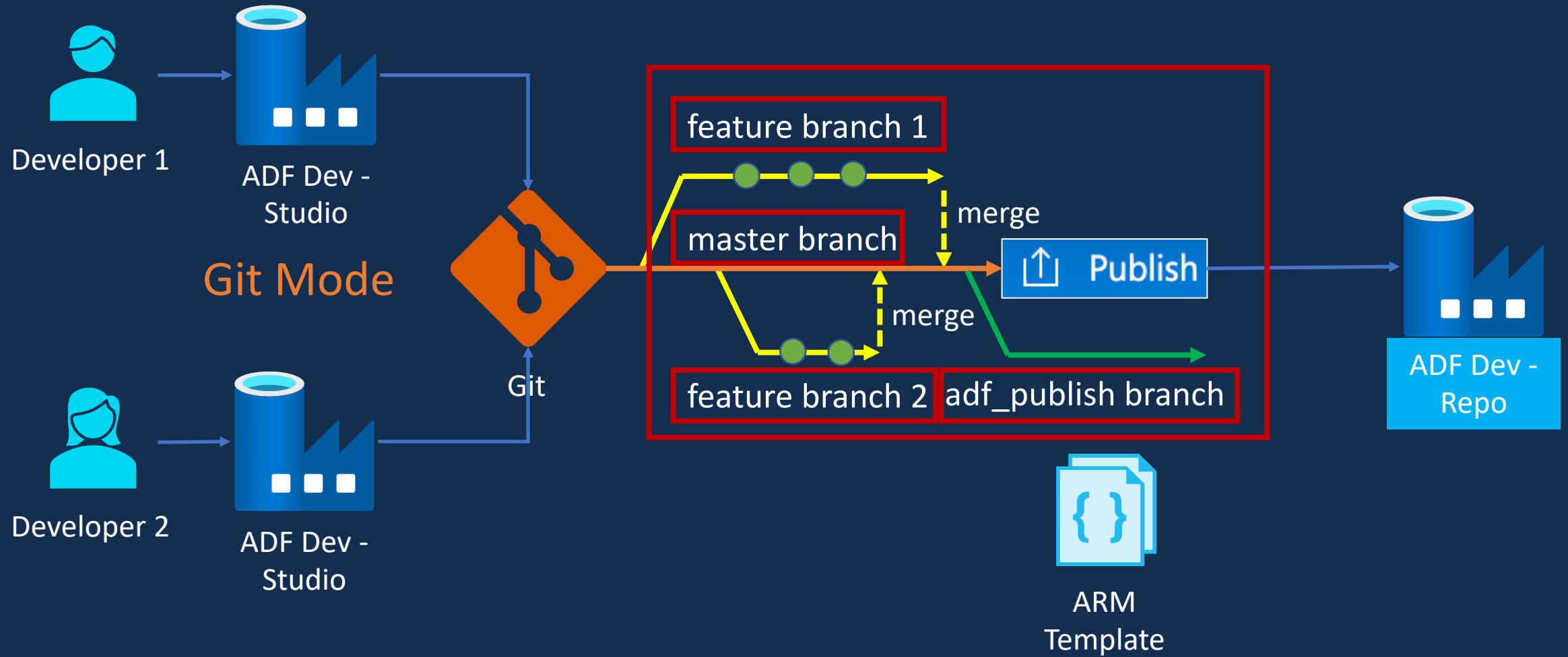
Env	Data Factory Name	Resource Group Name	GIT Enabled
dev	dev-ci-cd-demo-adf	dev-ci-cd-demo-rg	Y
test	test-ci-cd-demo-adf	test-ci-cd-demo-rg	N
prod	prod-ci-cd-demo-adf	prod-ci-cd-demo-rg	N



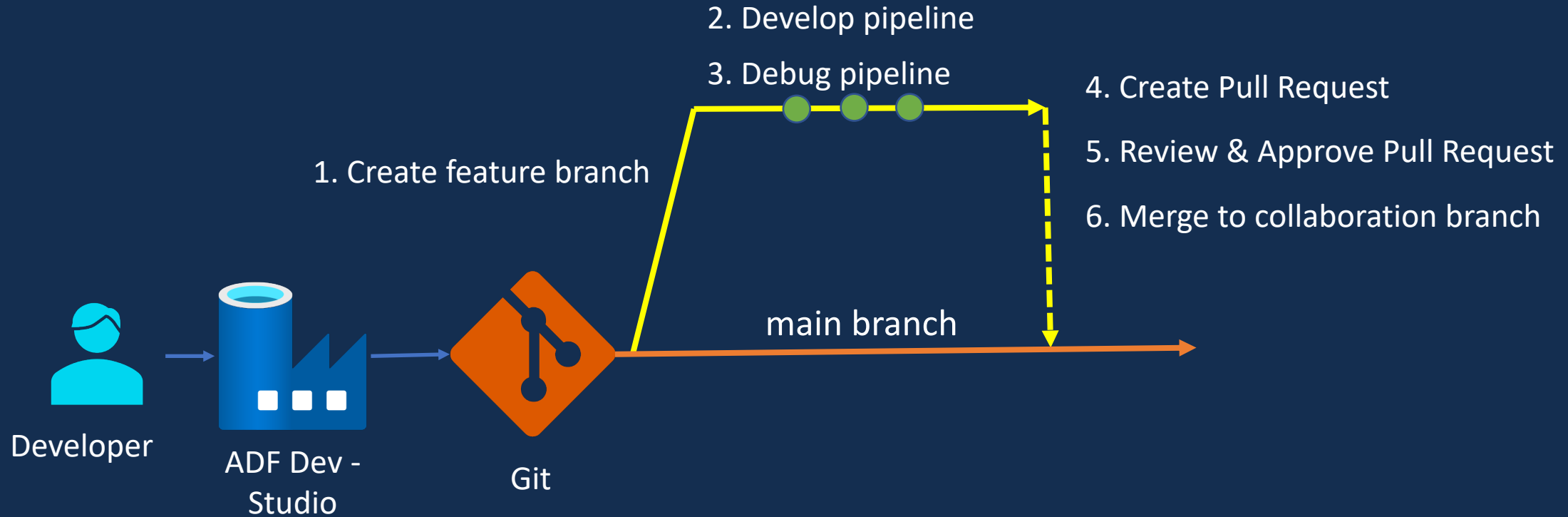
# ADF Git Configuration



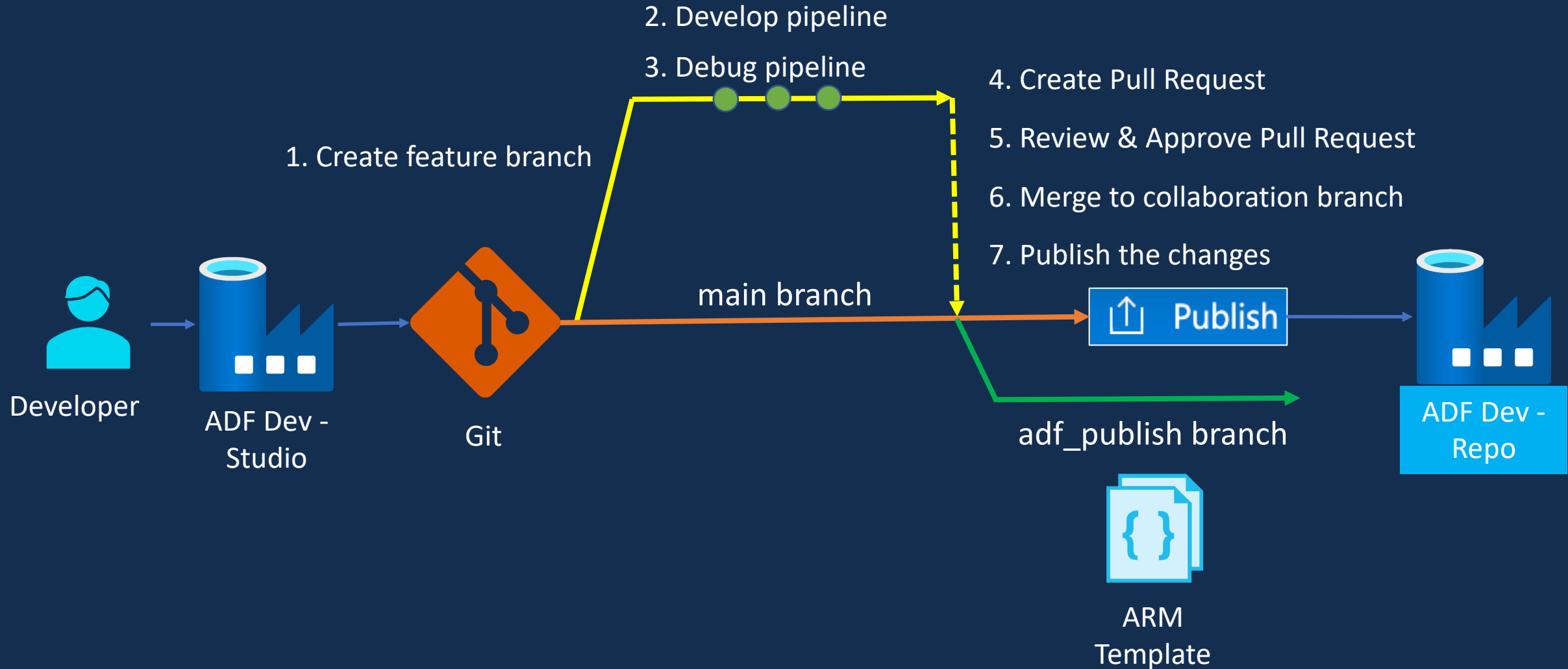
# Git Configuration



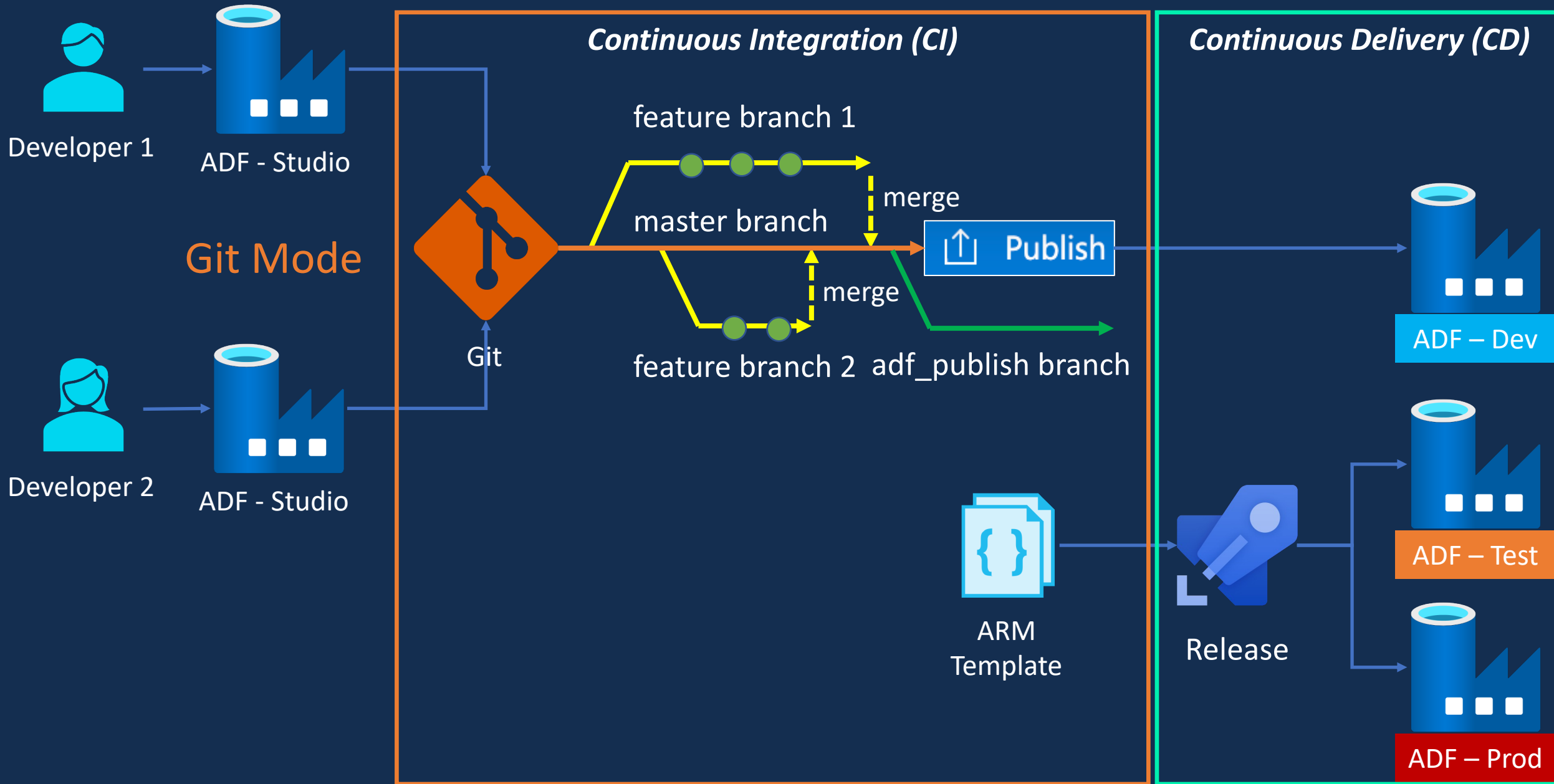
# Continuous Integration - Code



# Continuous Integration - Build

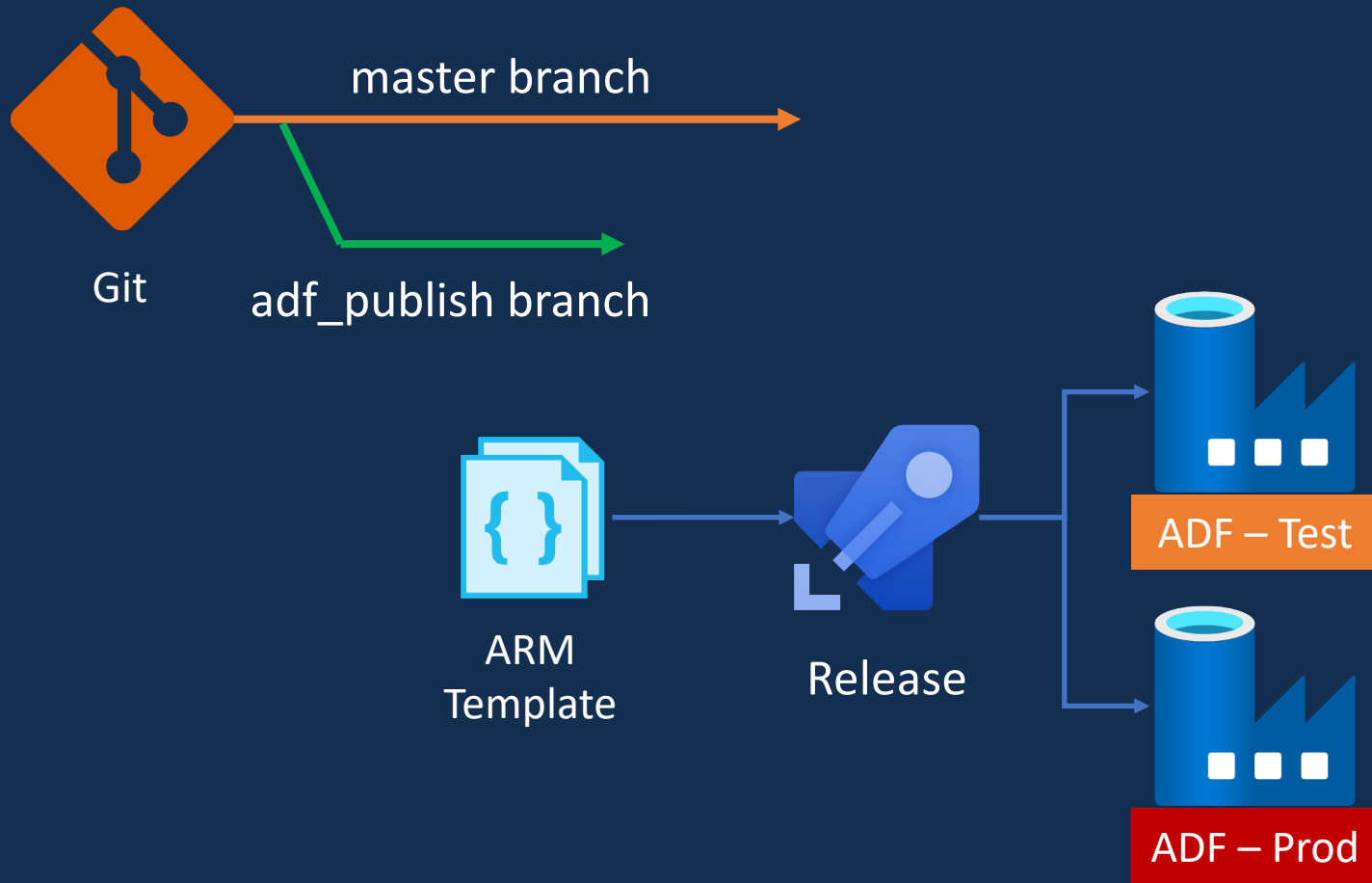


# Continuous Delivery

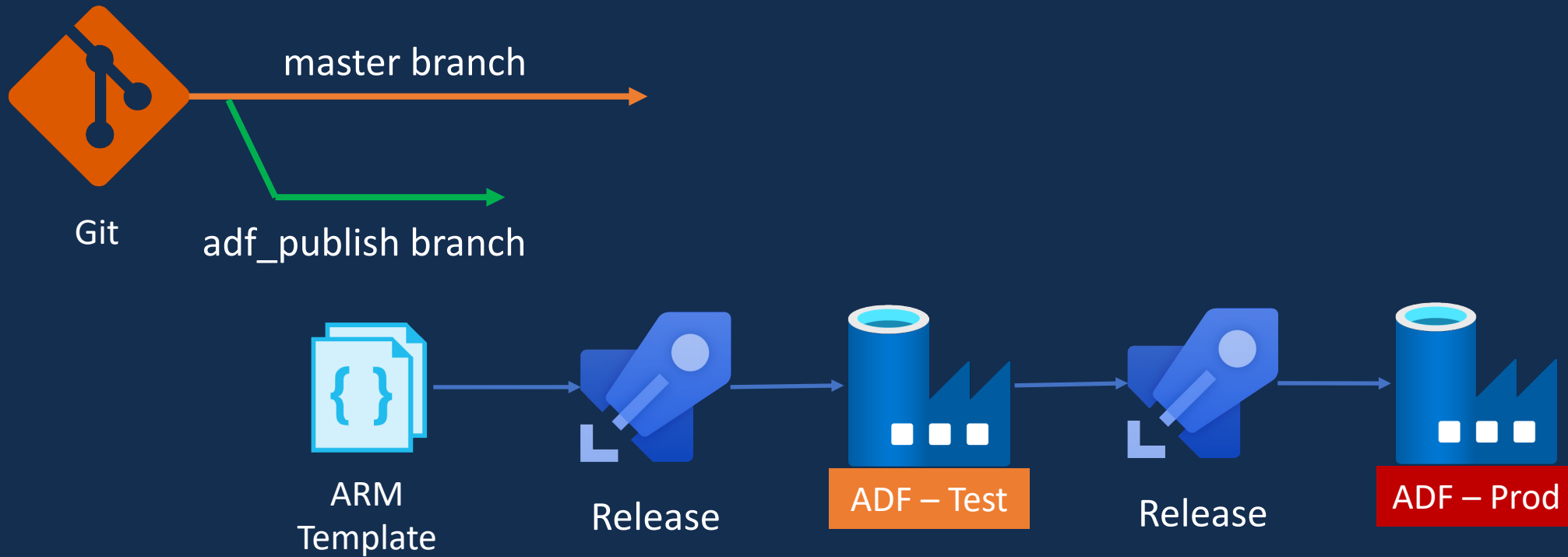




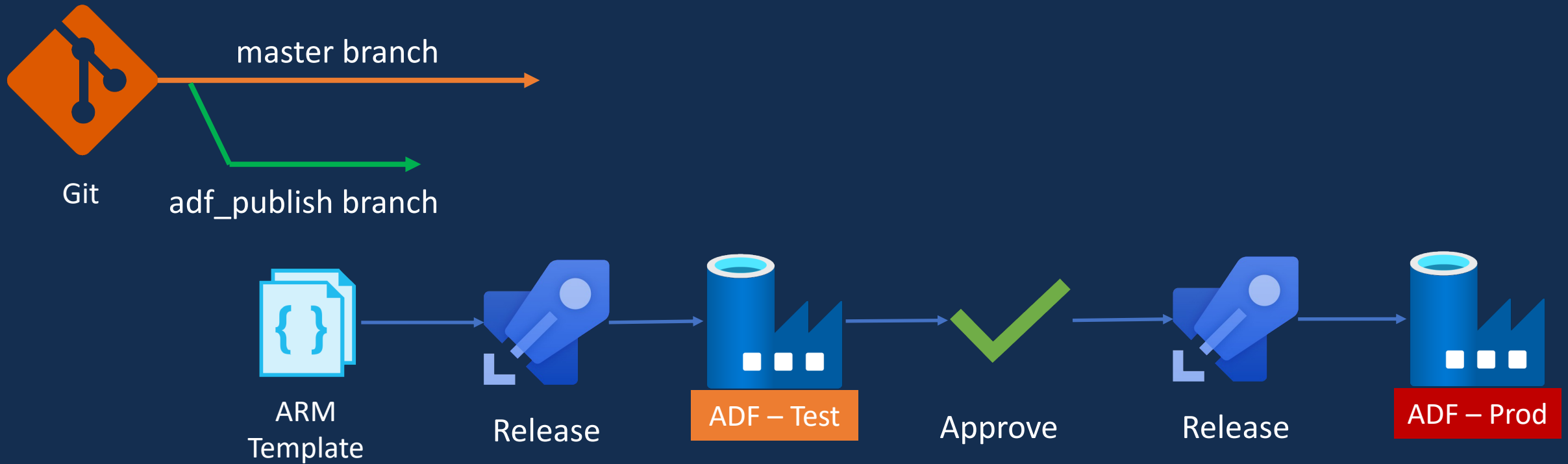
# Continuous Delivery



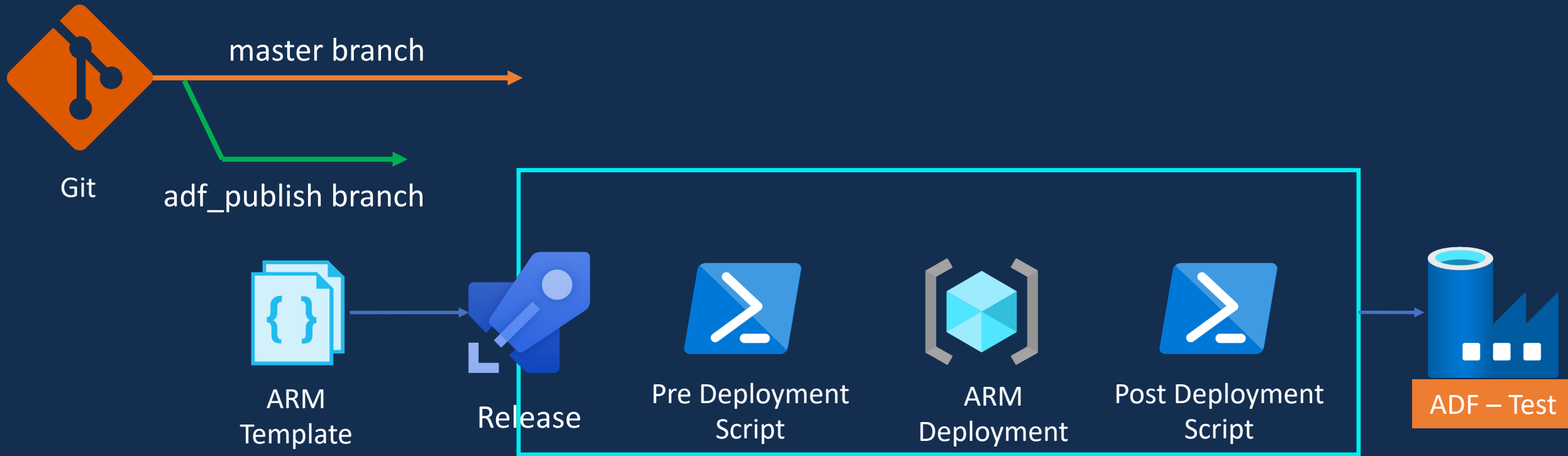
# Continuous Delivery



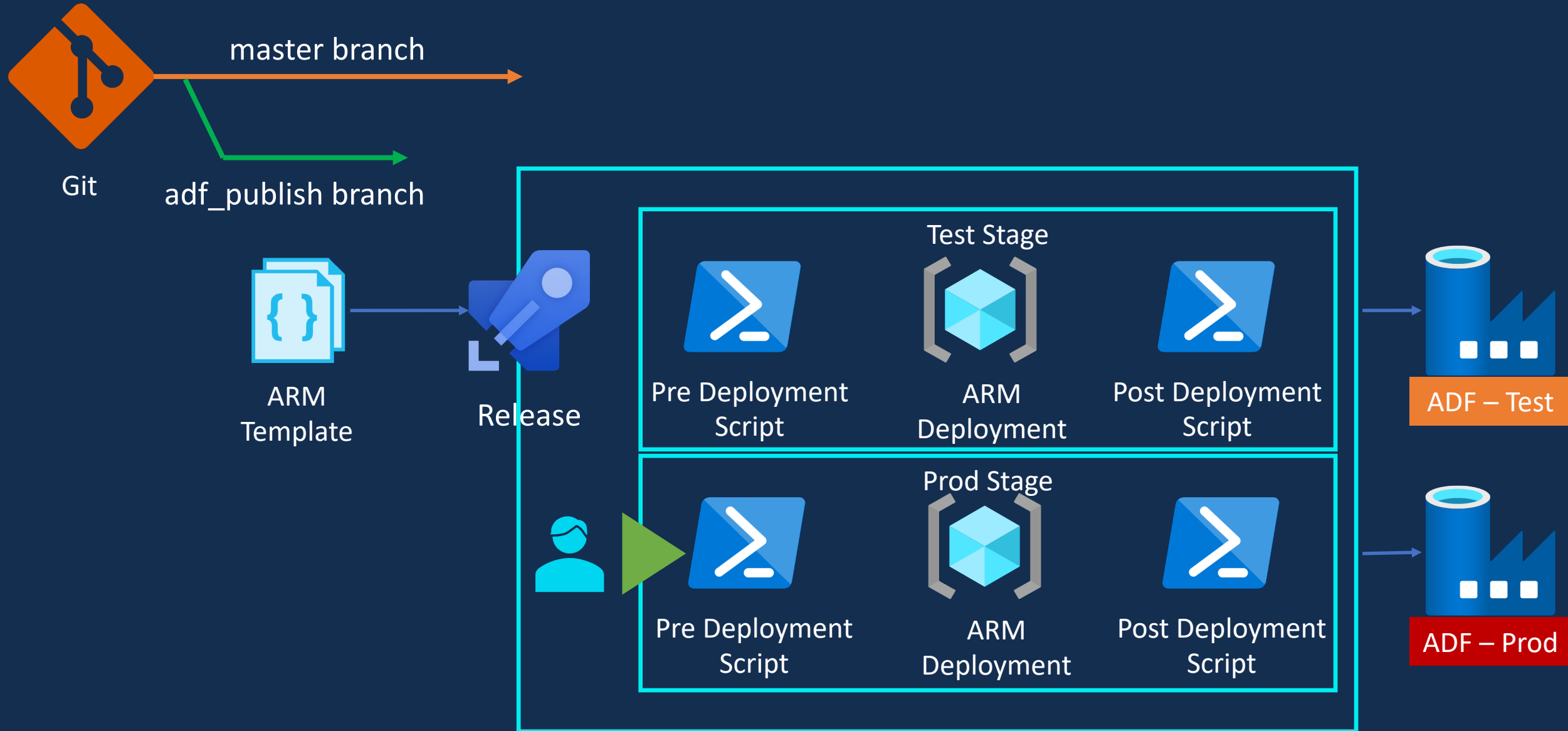
# Continuous Delivery



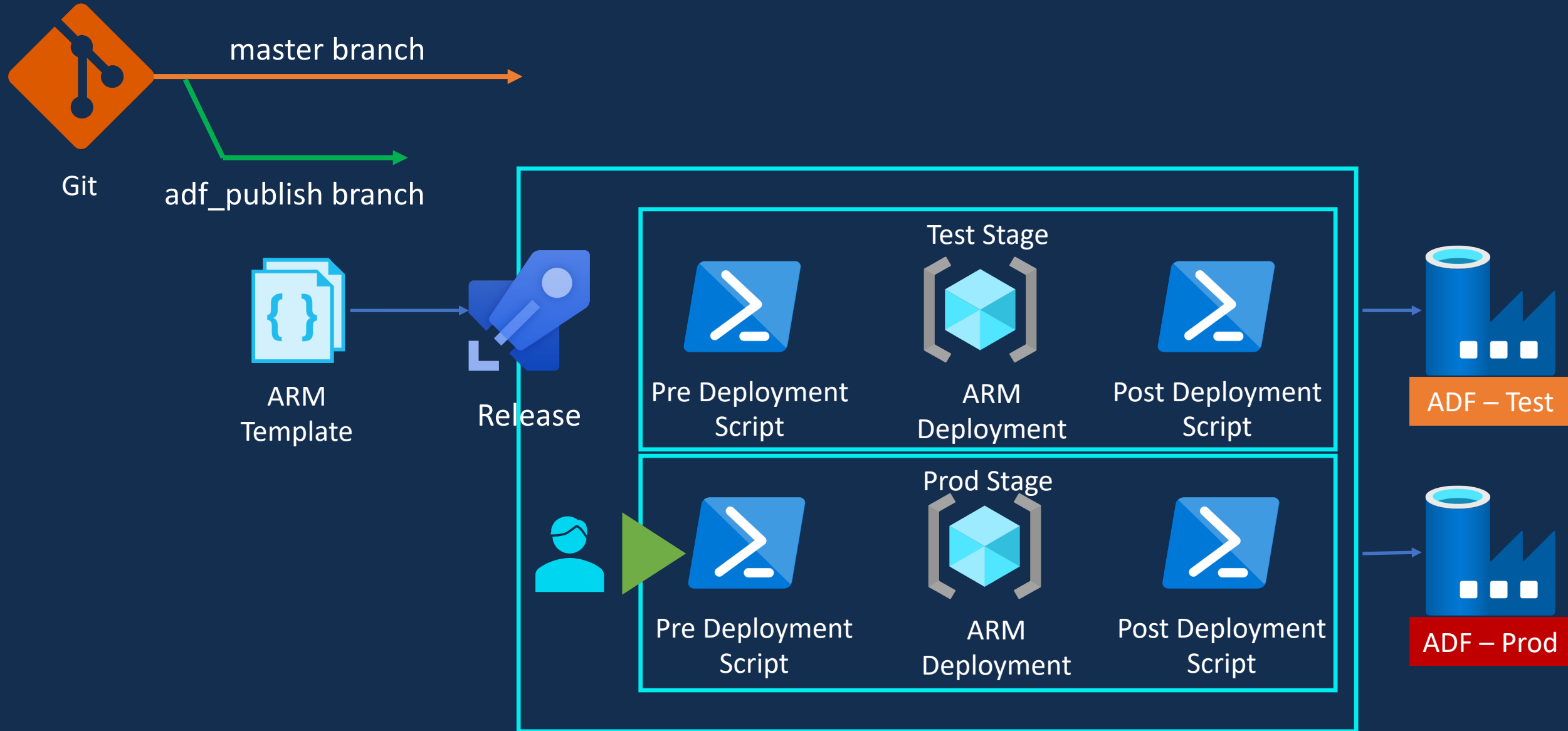
# Continuous Delivery – Release Pipeline



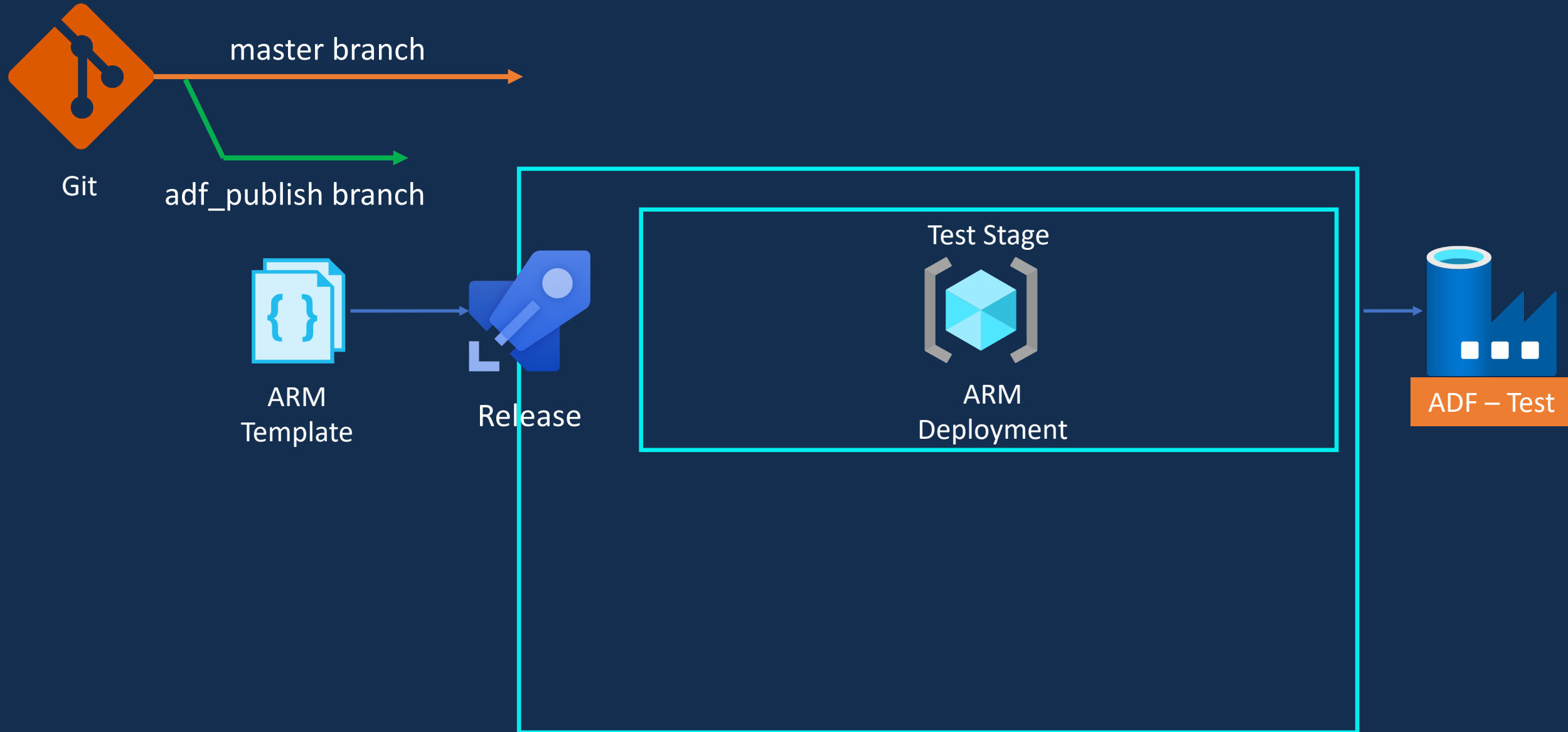
# Continuous Delivery – Release Pipeline



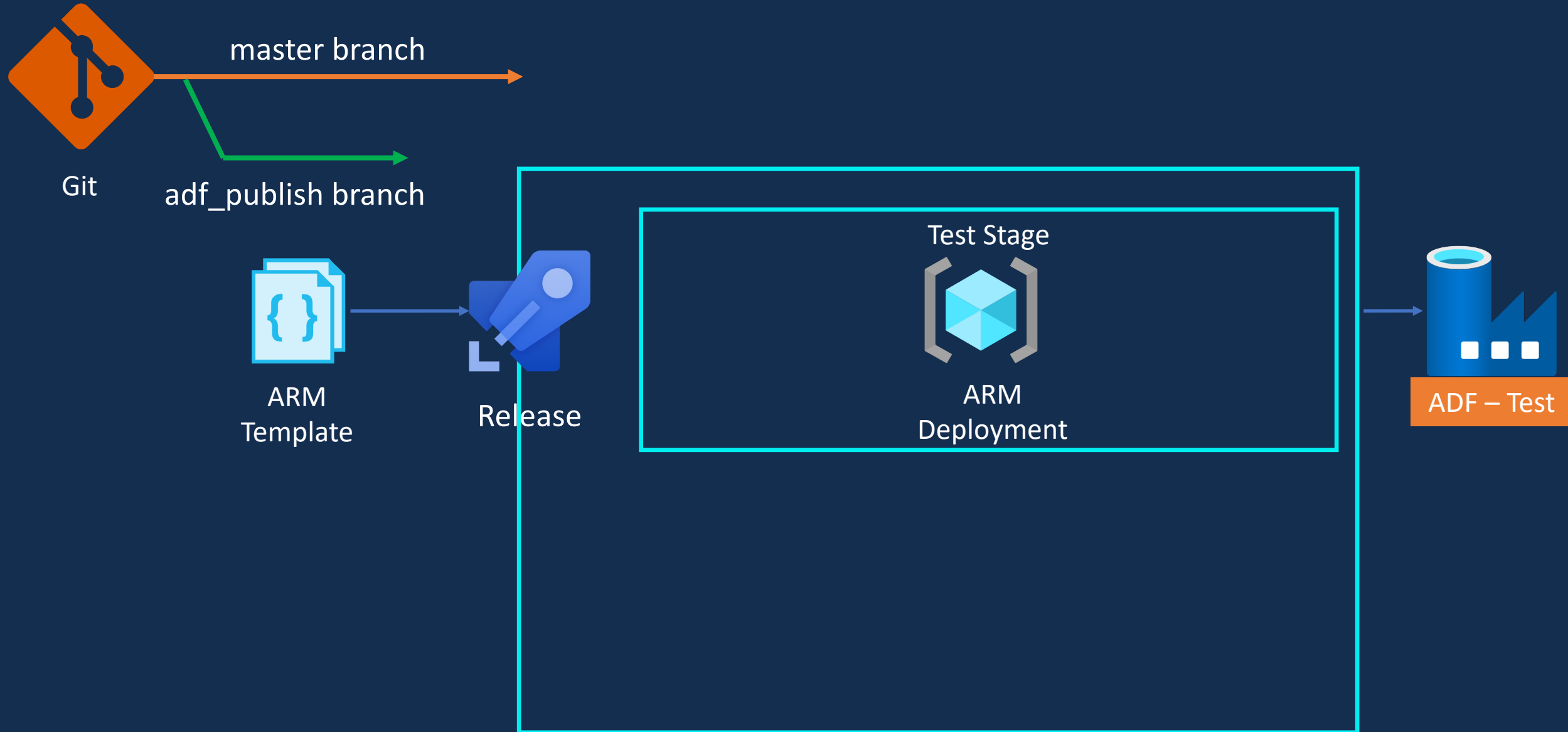
# Continuous Delivery – Release Pipeline



# Continuous Delivery – Release Pipeline

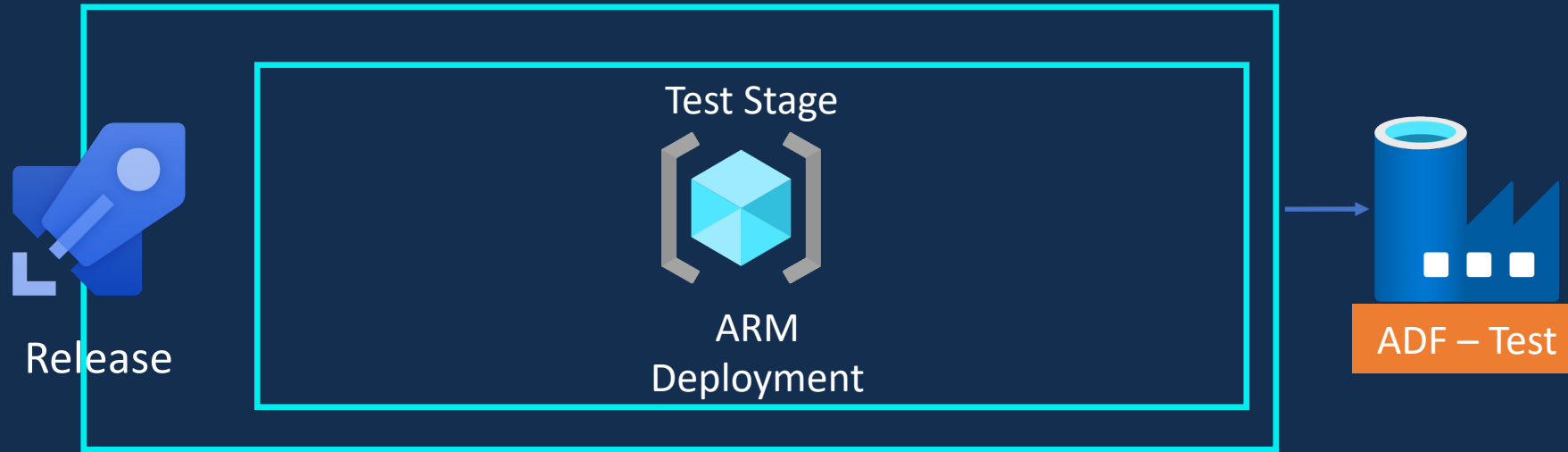


# Release Pipeline - Issues





# Release Pipeline - Issues

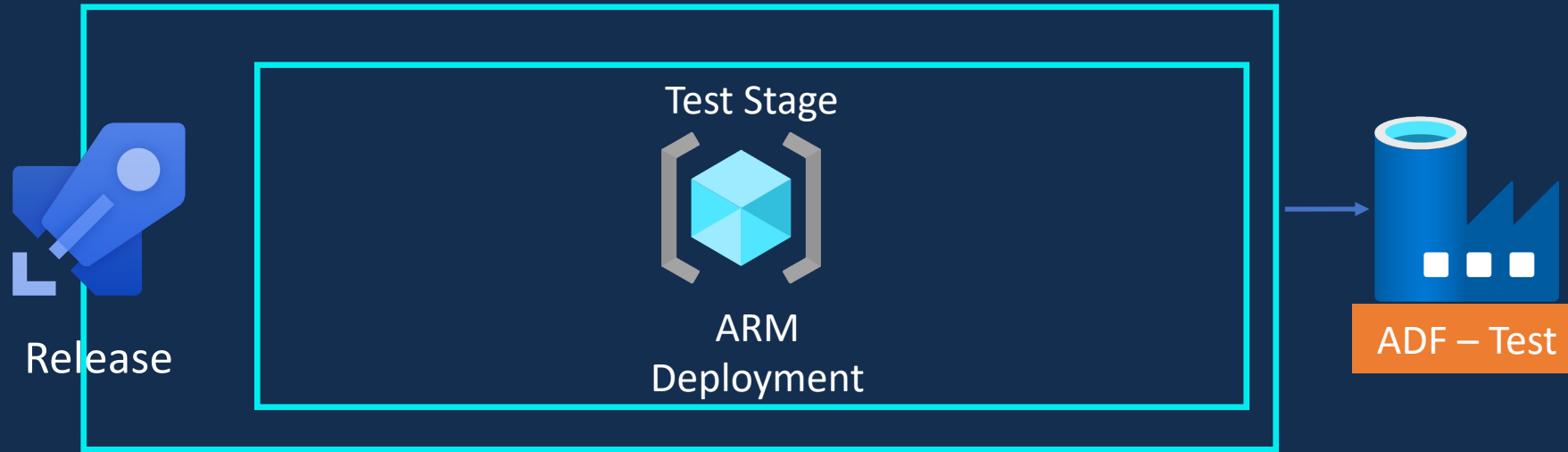


Deleted Objects



Active Triggers

# Release Pipeline - Issues

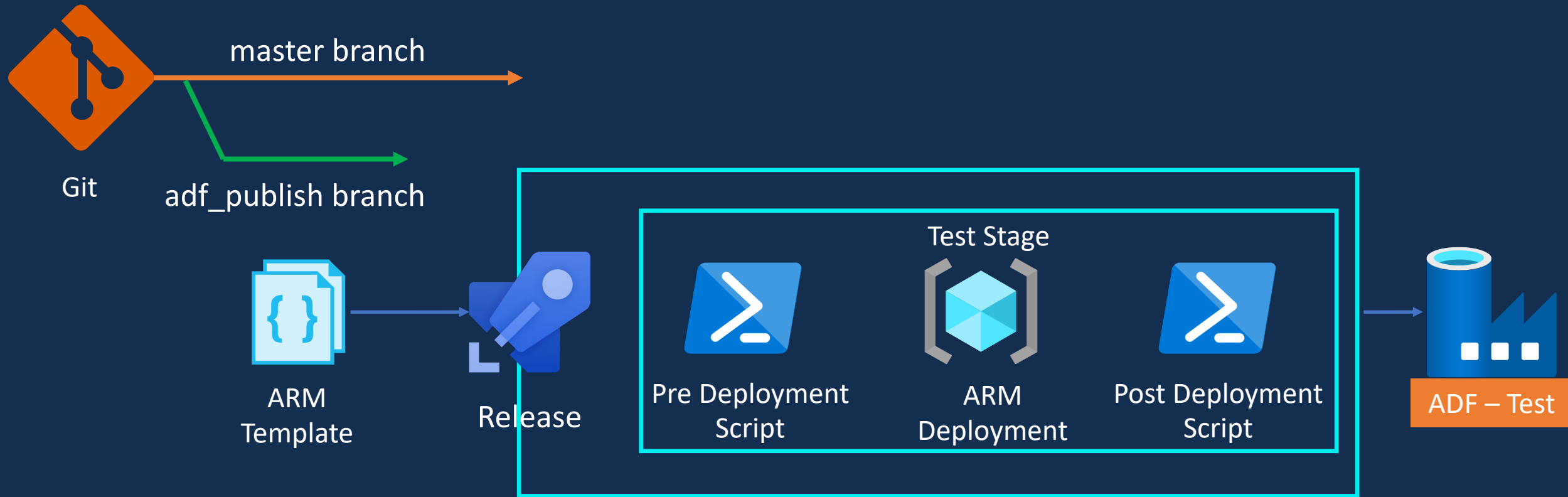


Deleted Objects

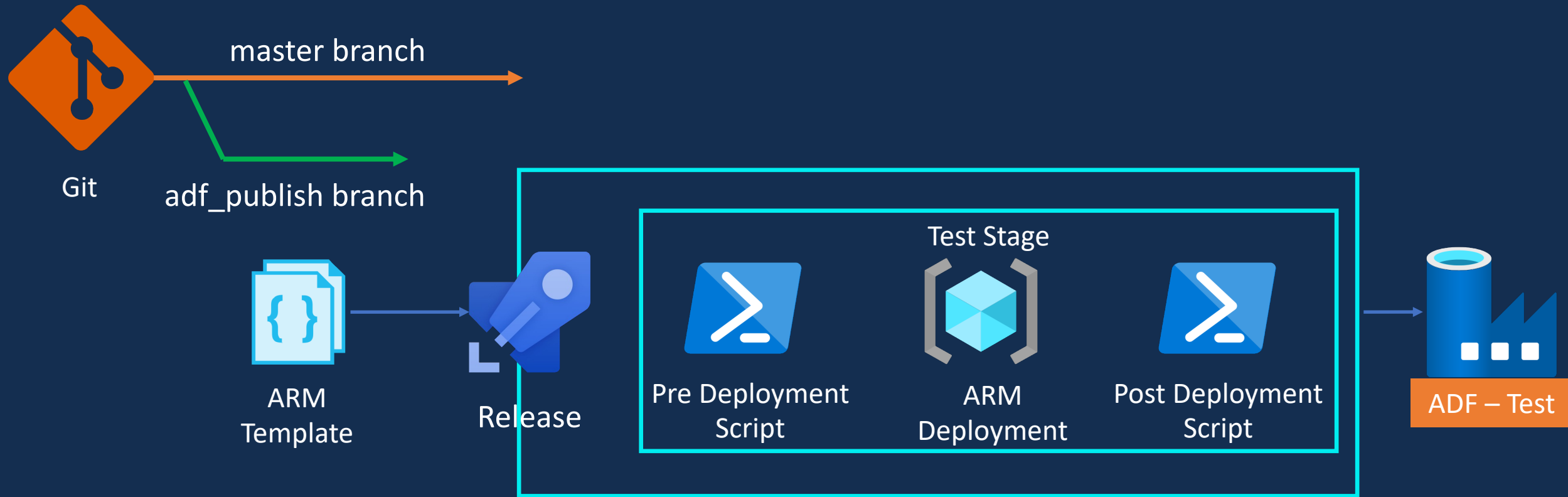


Active Triggers

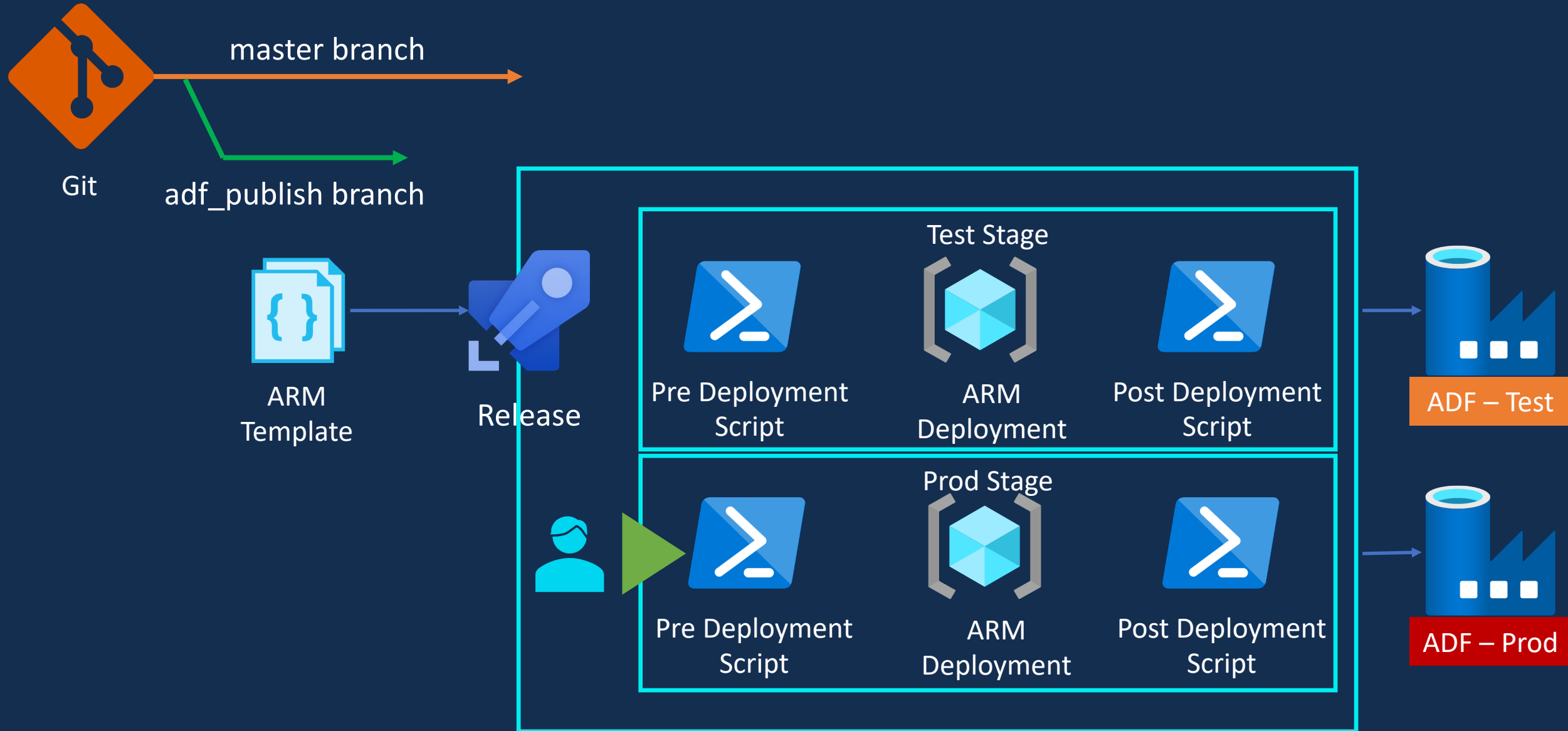
# Release Pipeline – Pre & Post Deployment



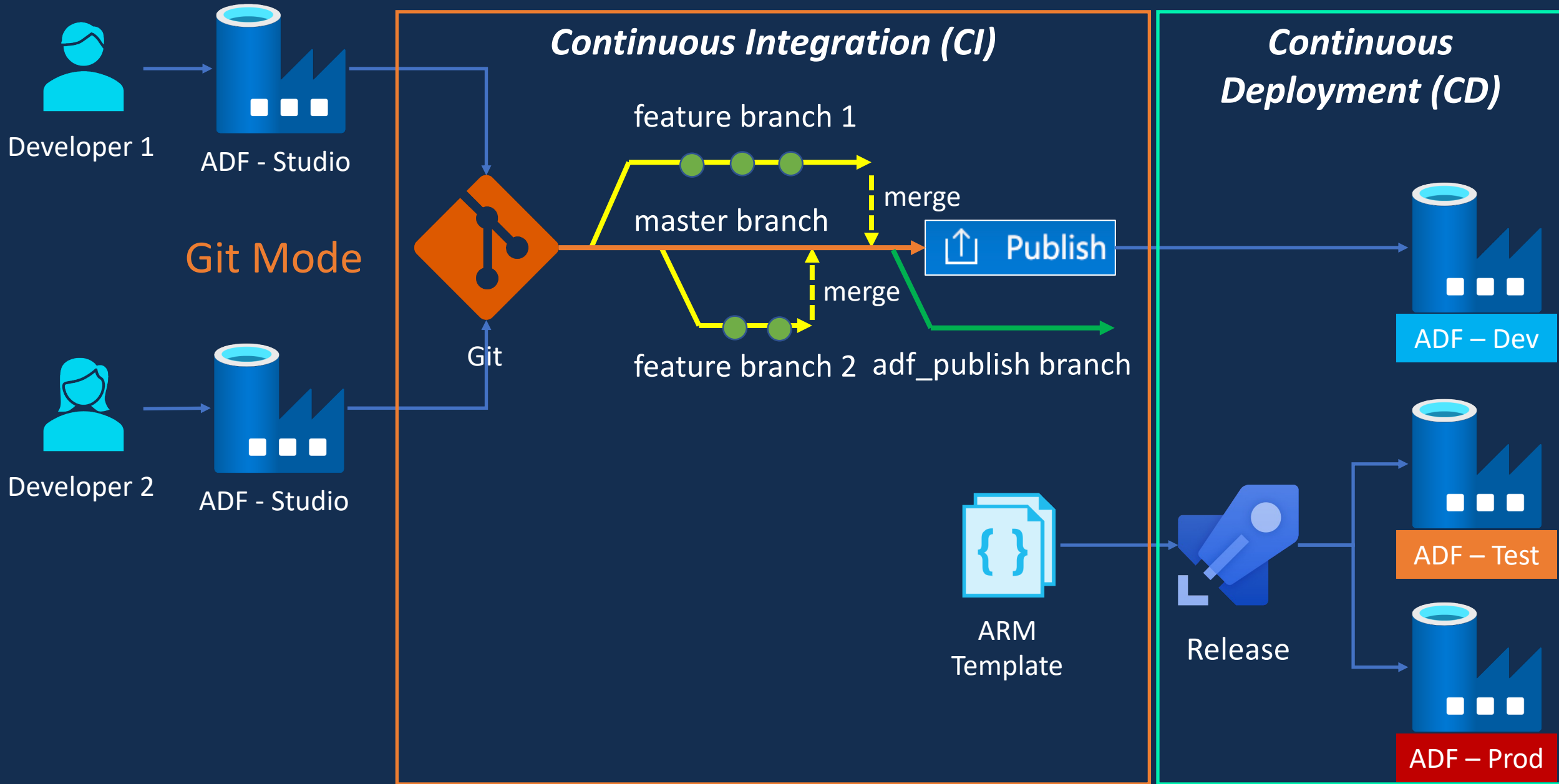
# Release Pipeline – Deployment to Test



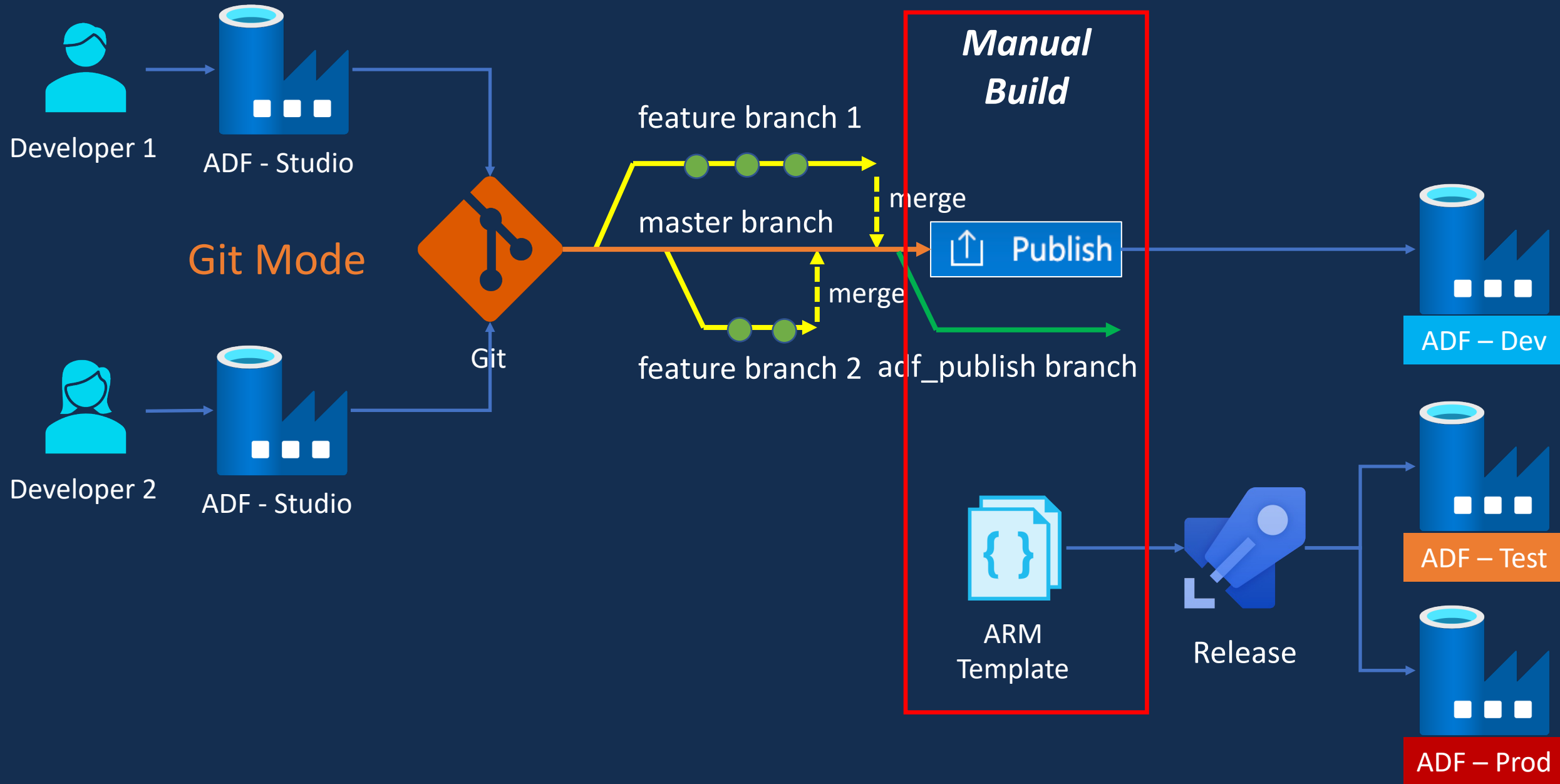
# Release Pipeline – Deployment to Prod



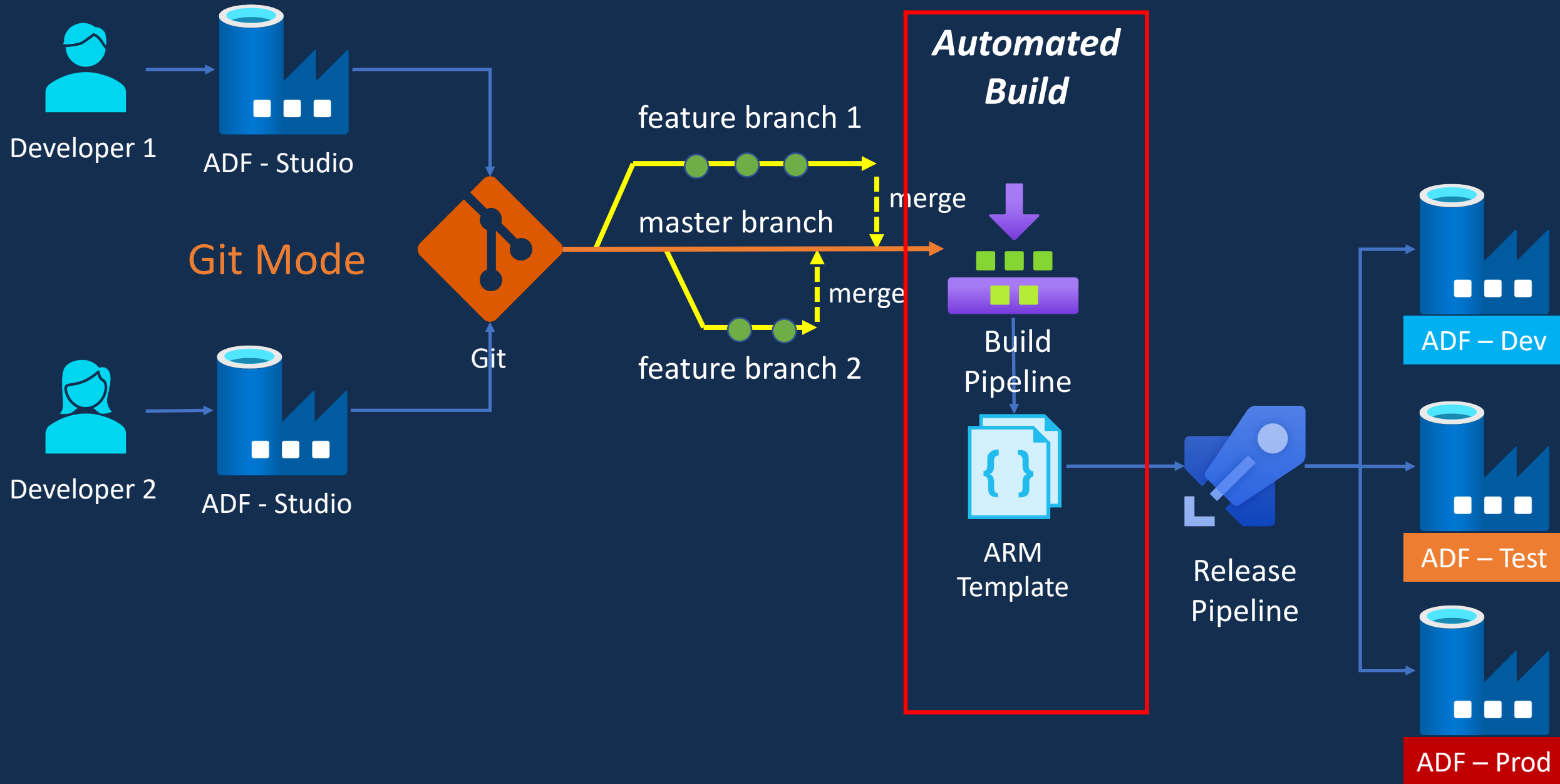
# CI/CD Option 1 – Using ADF Publish



# CI/CD Option 1 – Using ADF Publish

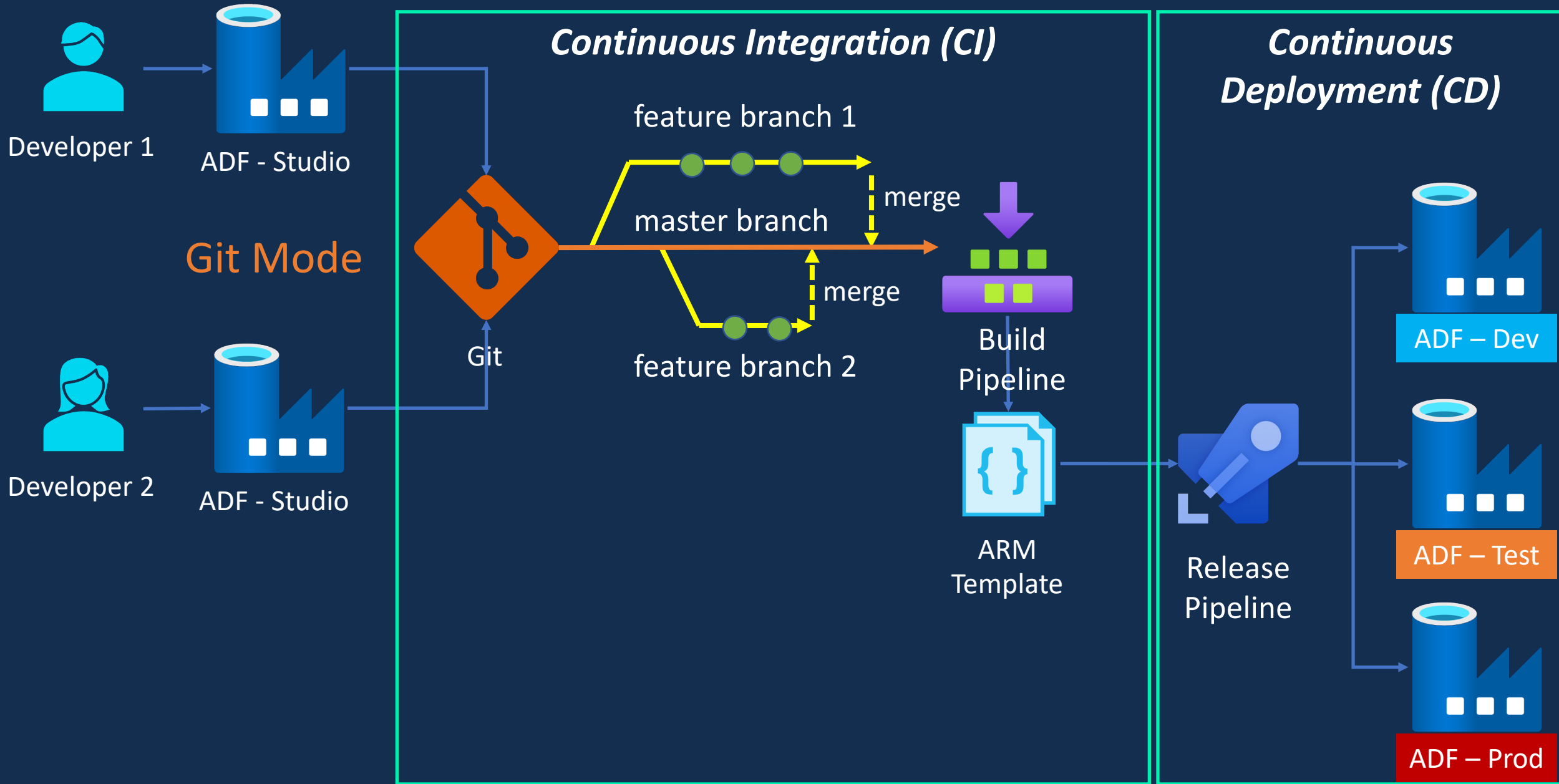


# CI/CD Option 2 – Using Build Pipeline

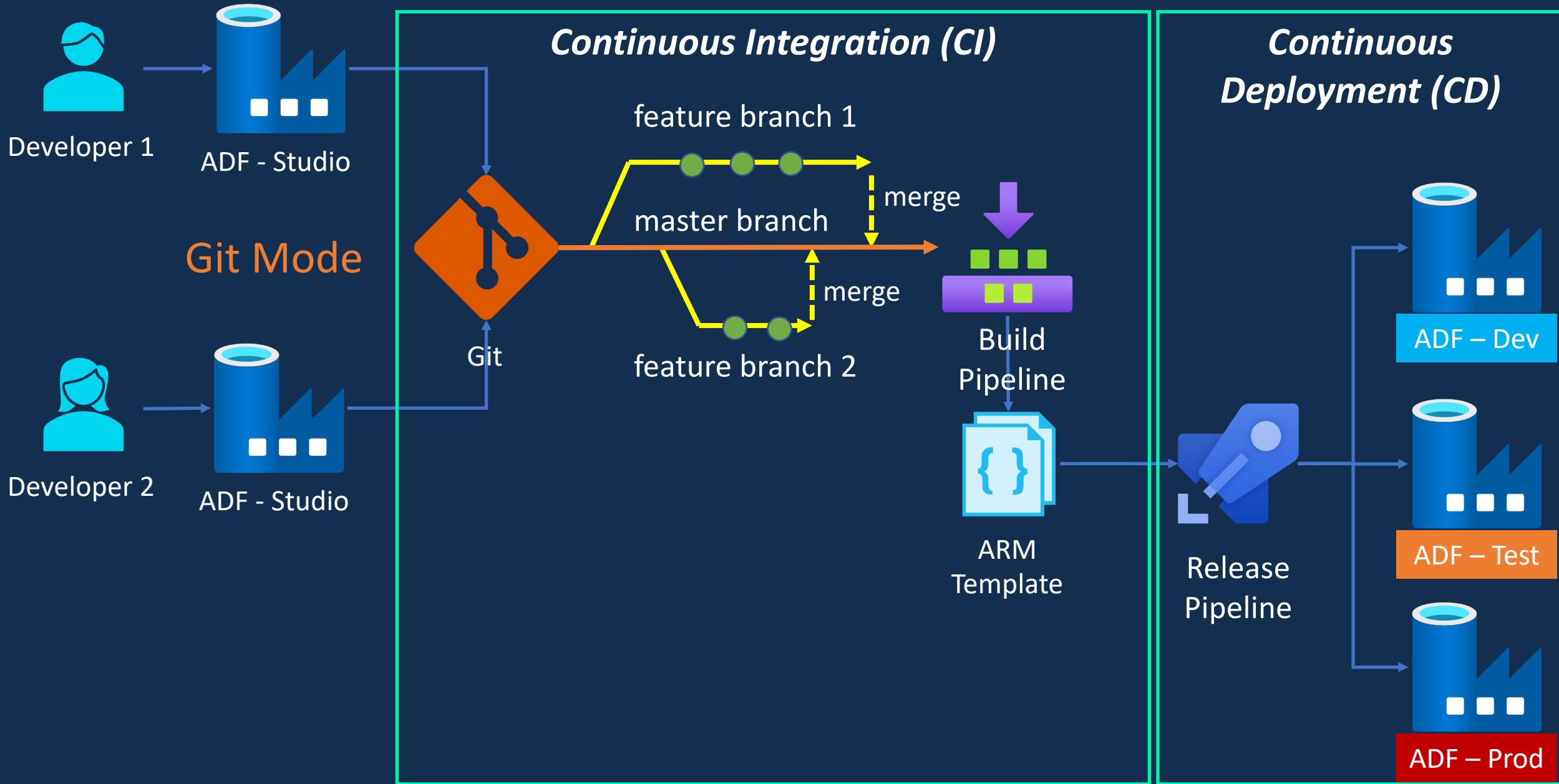




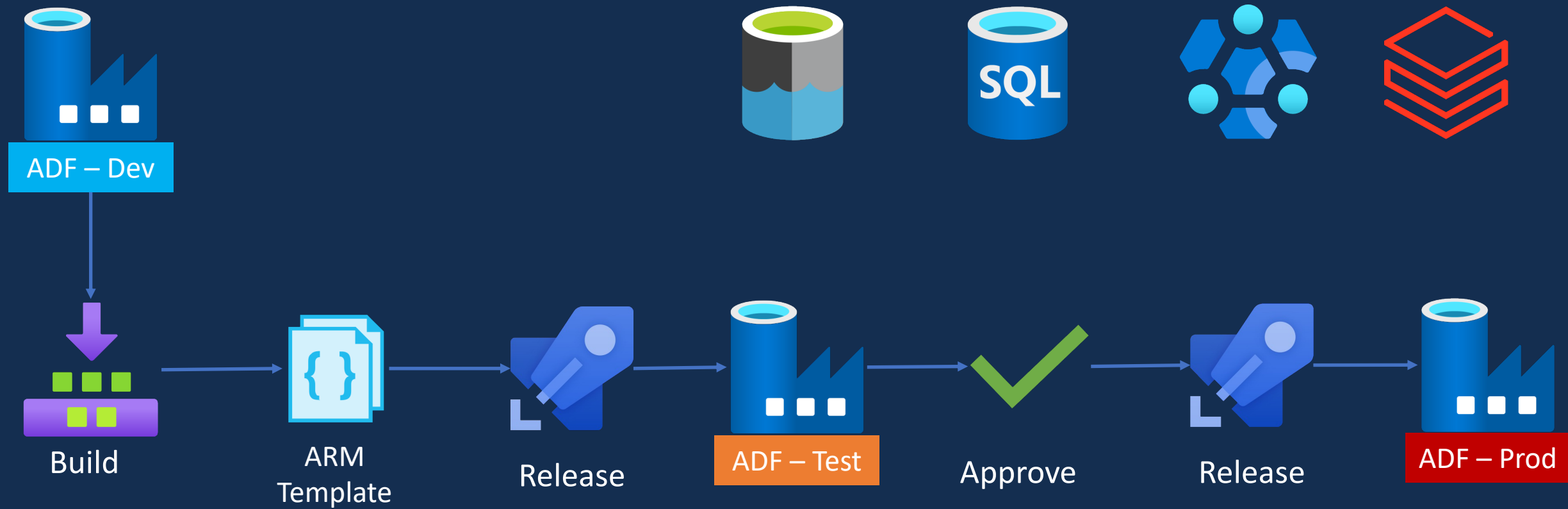
# CI/CD Option 2 – Using Build Pipeline



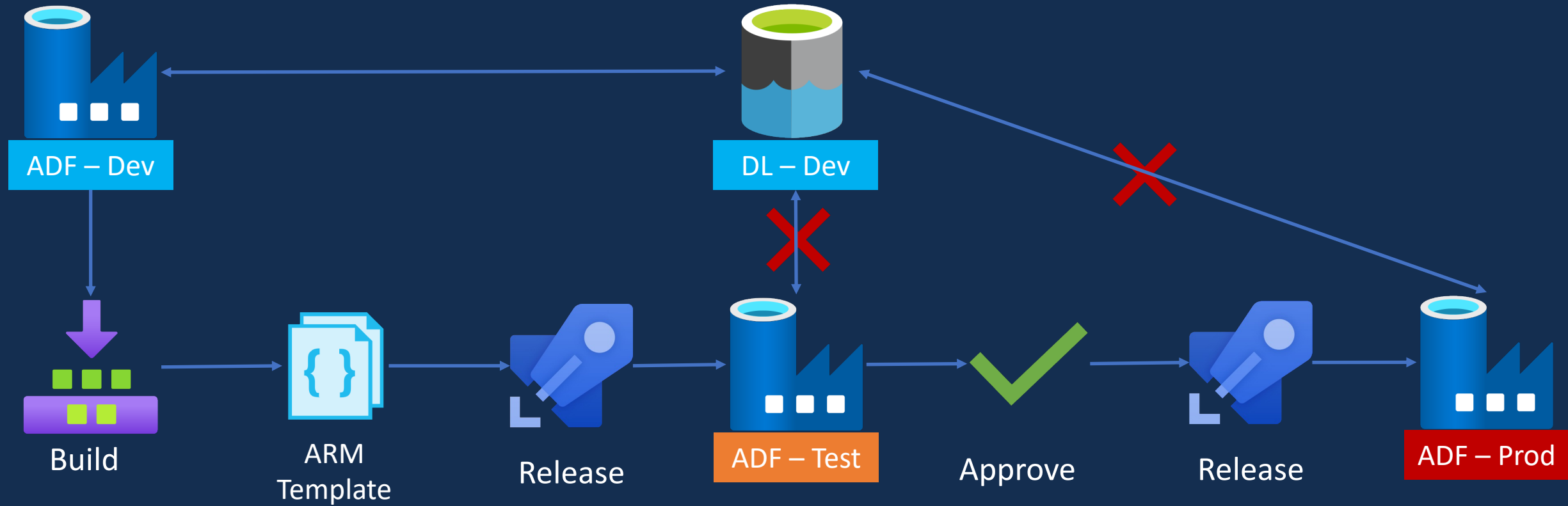
# Continuous Integration/ Continuous Delivery



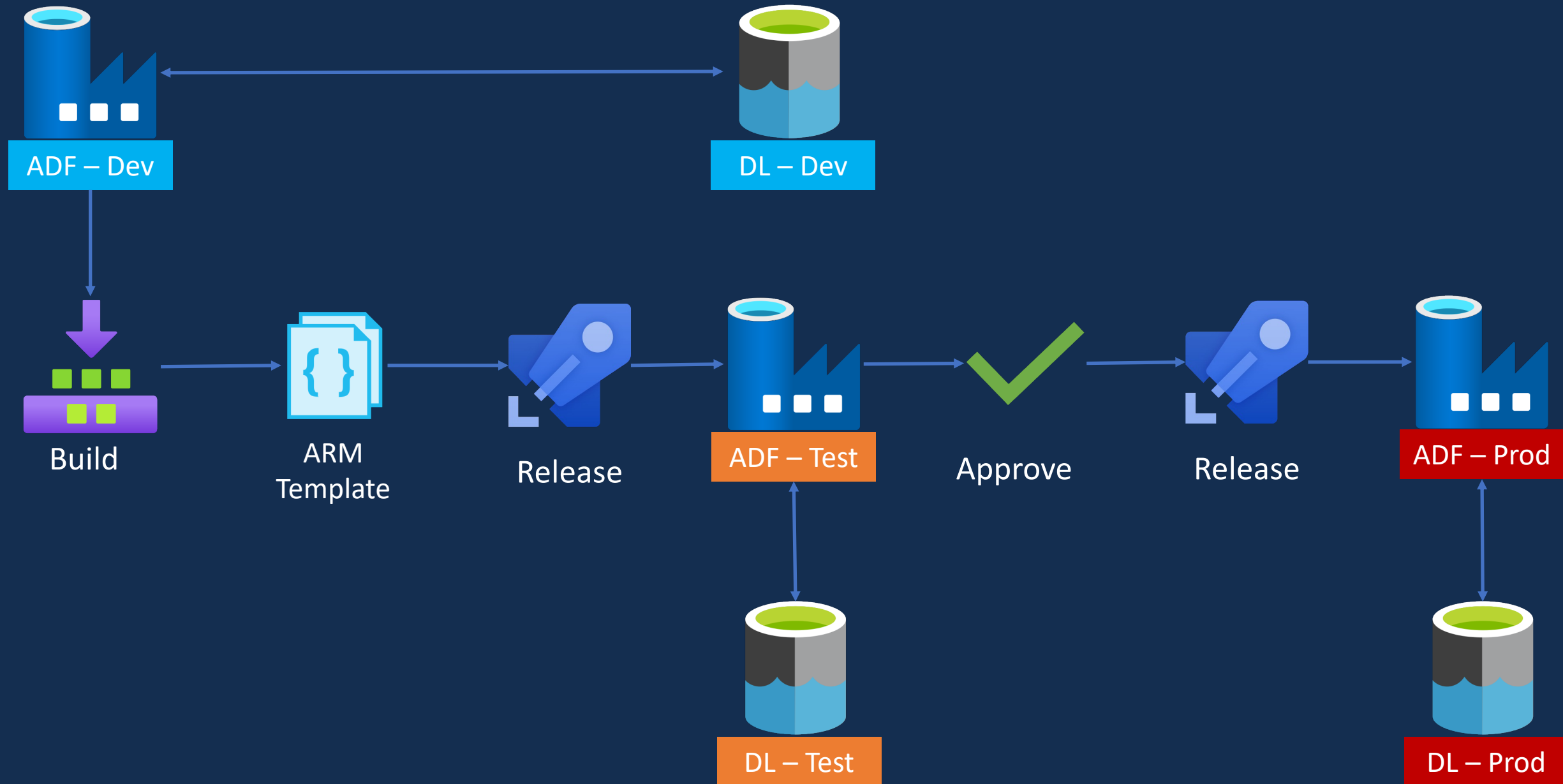
# CI/CD Scenario – Data Lake Access



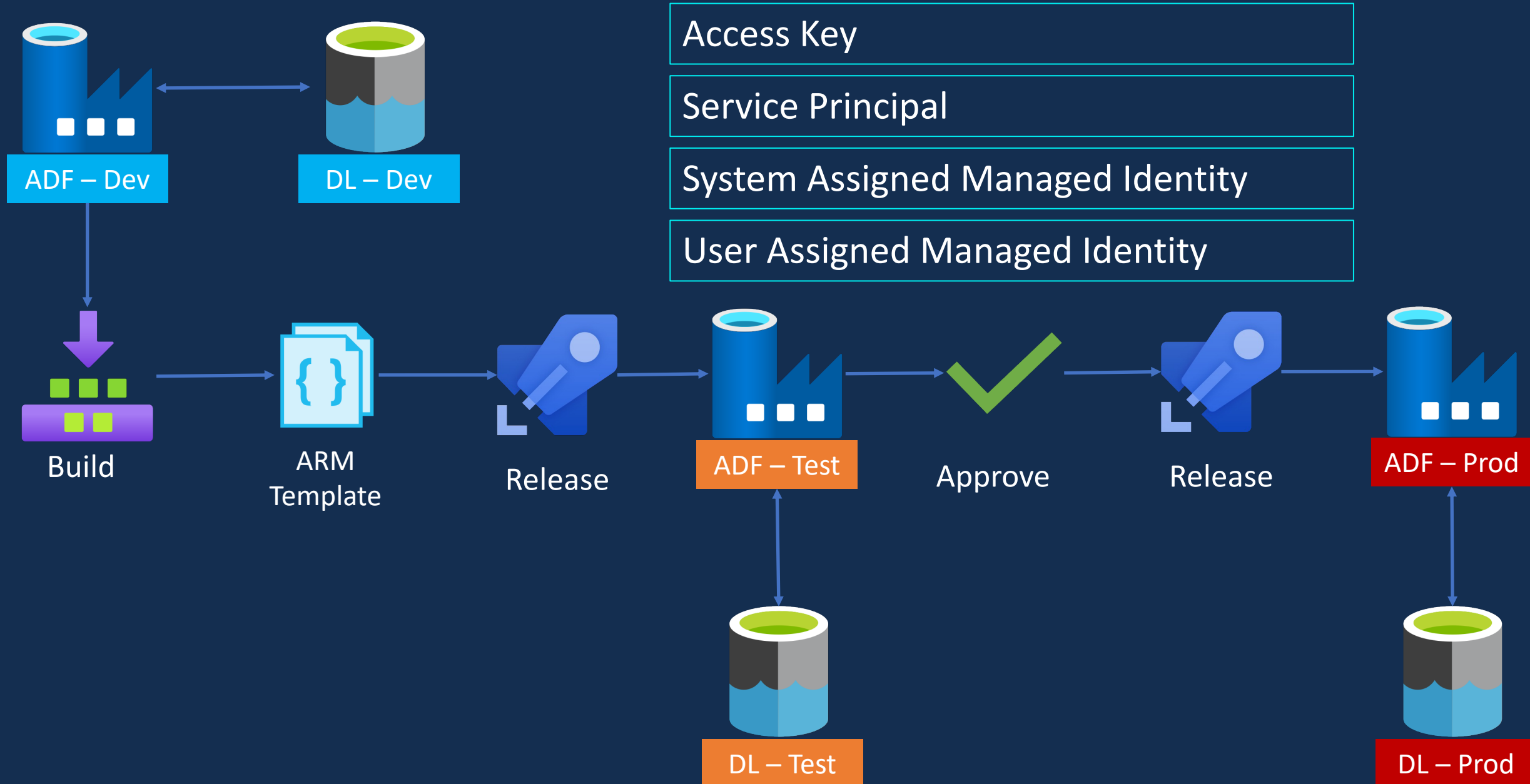
# CI/CD Scenario – Data Lake Access



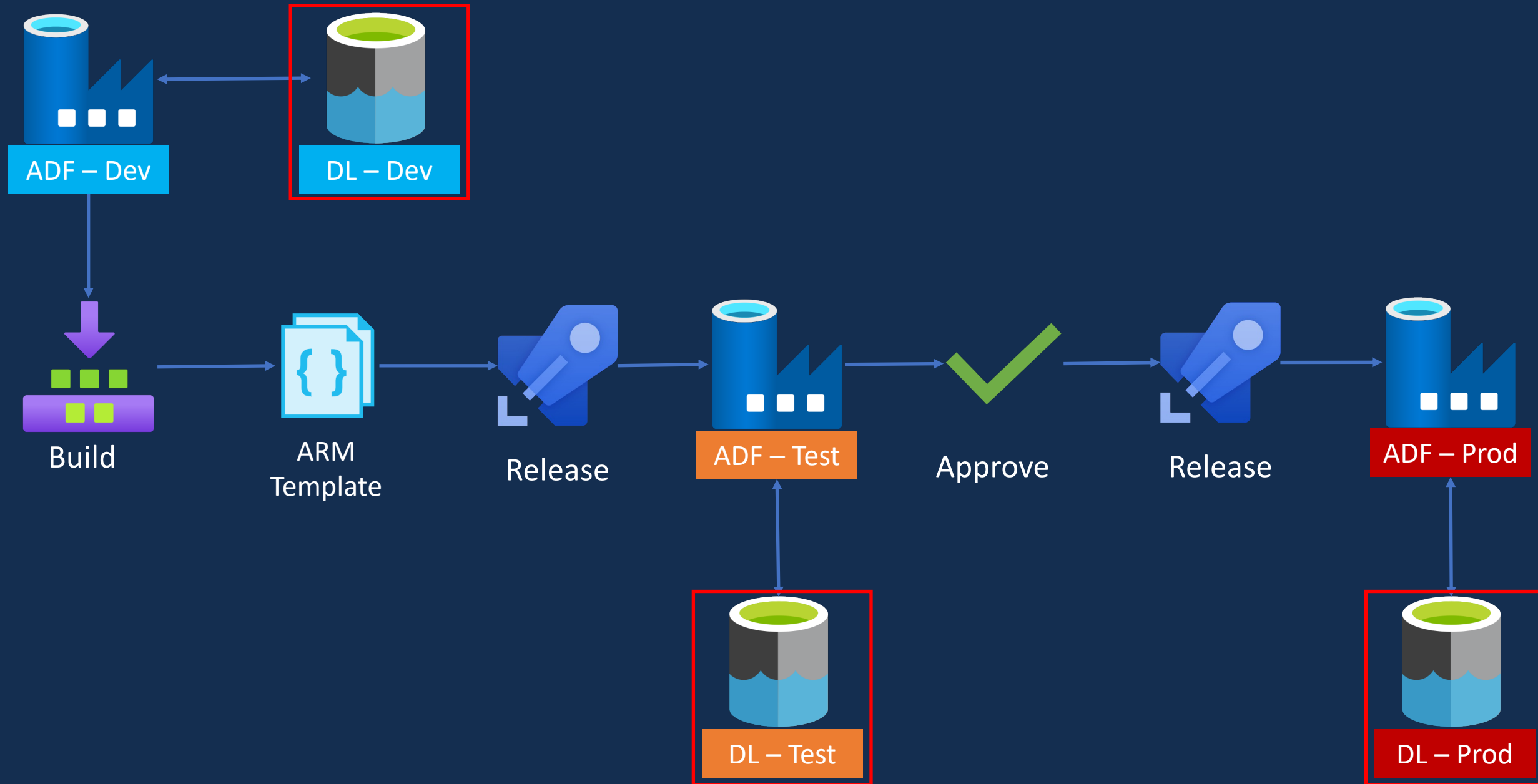
# CI/CD Scenario – Data Lake Access



# CI/CD Scenario – Data Lake Access



# CI/CD Scenario – Data Lake Access

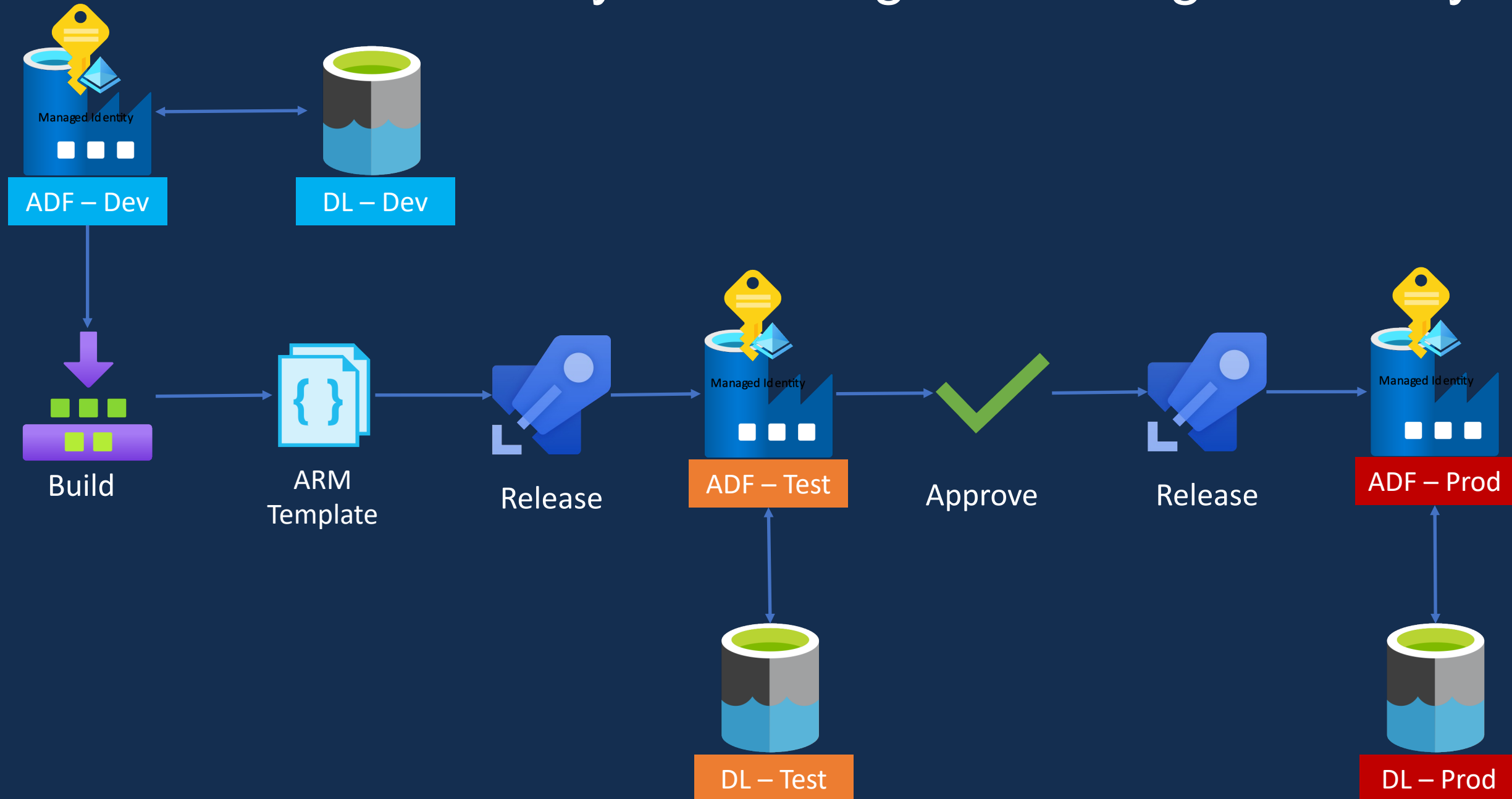


# Data Lake Storage Set-up

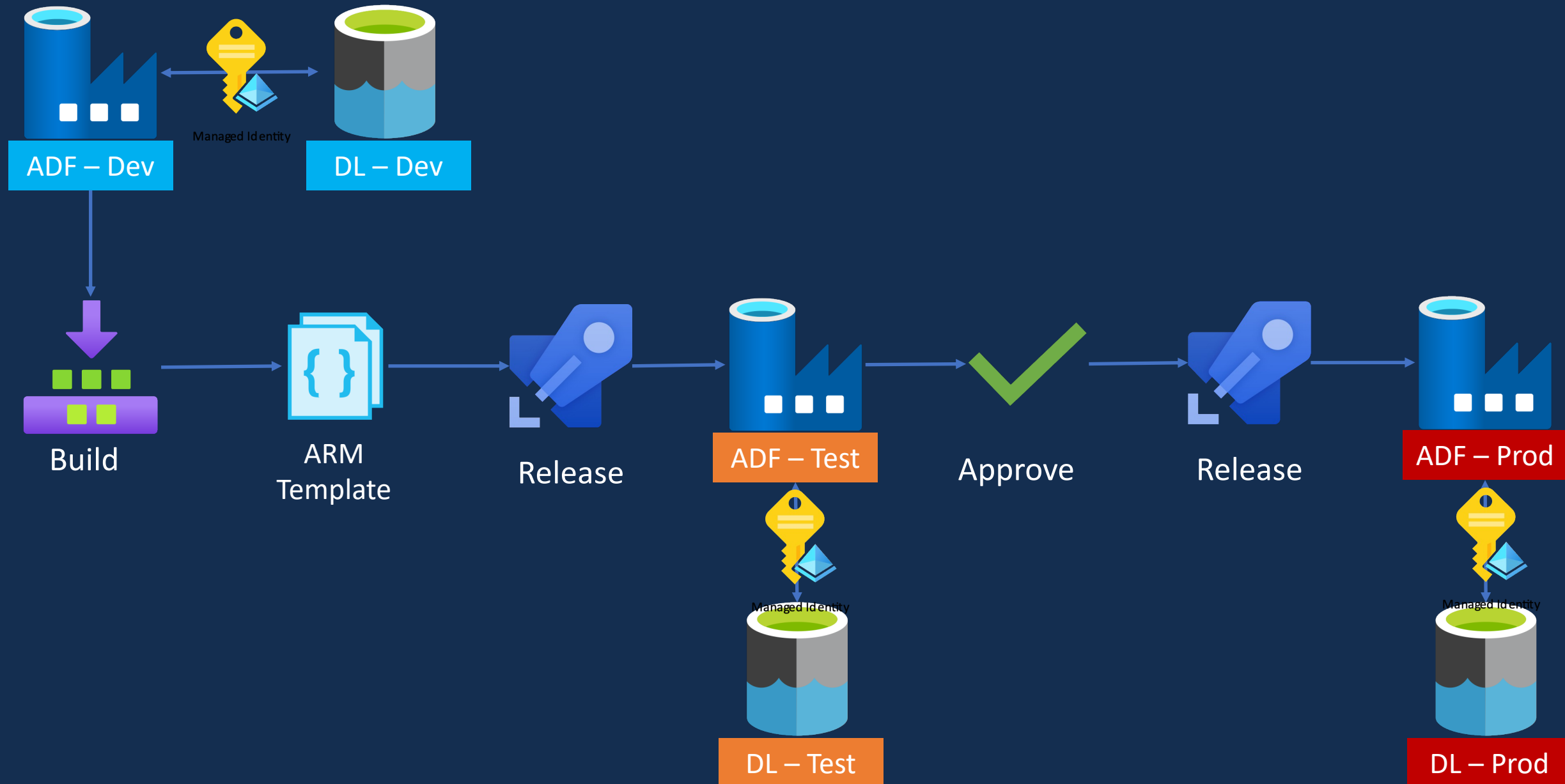
Env	Data Factory Name	Resource Group Name	Data Lake Name	GIT Enabled
dev	dev-ci-cd-demo-adf	dev-ci-cd-demo-rg	devcicddemodl	Y
test	test-ci-cd-demo-adf	test-ci-cd-demo-rg	testcicddemodl	N
prod	prod-ci-cd-demo-adf	prod-ci-cd-demo-rg	prodcicddemodl	N



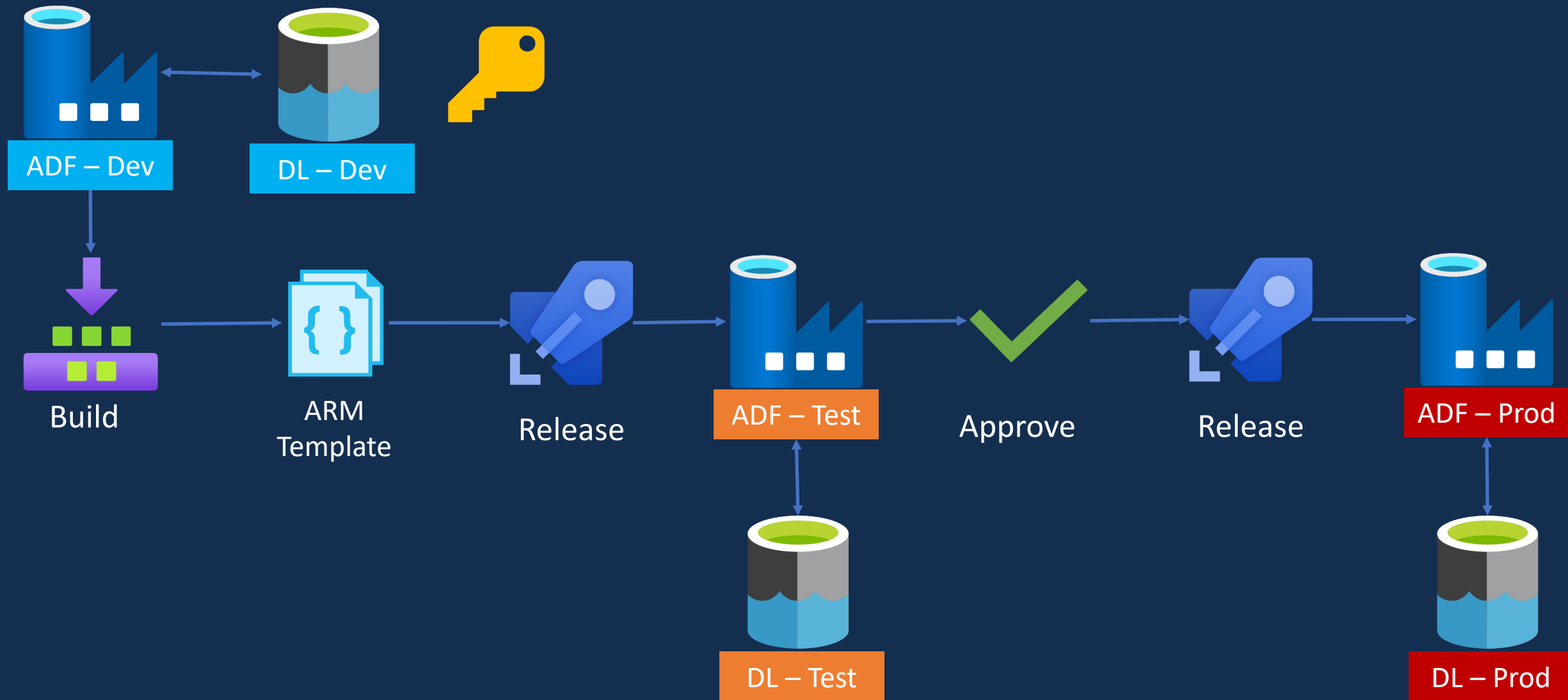
# Data Lake Access via System Assigned Managed Identity



# Data Lake Access via System Assigned Managed Identity



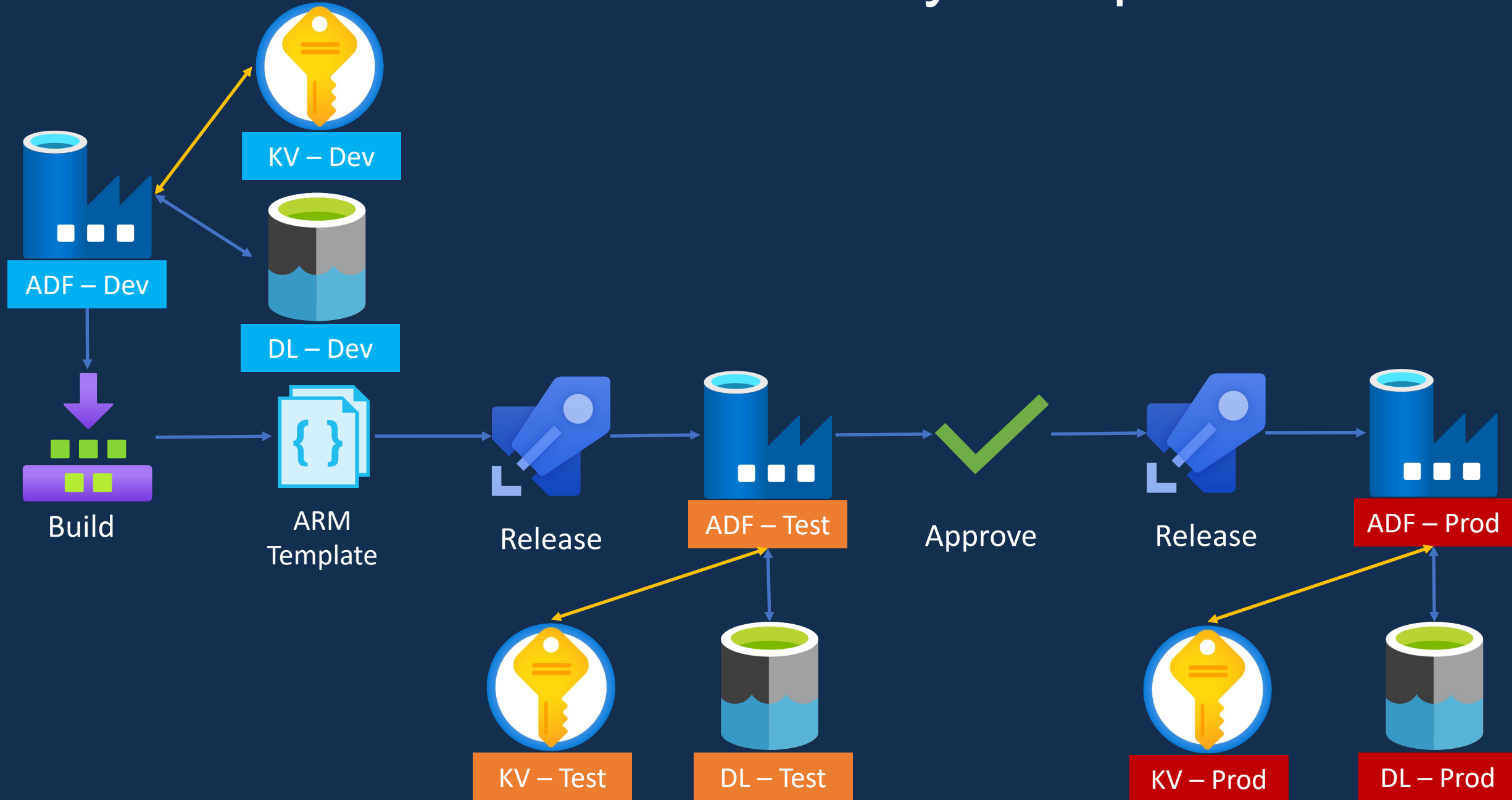
# Data Lake Access via Access Keys



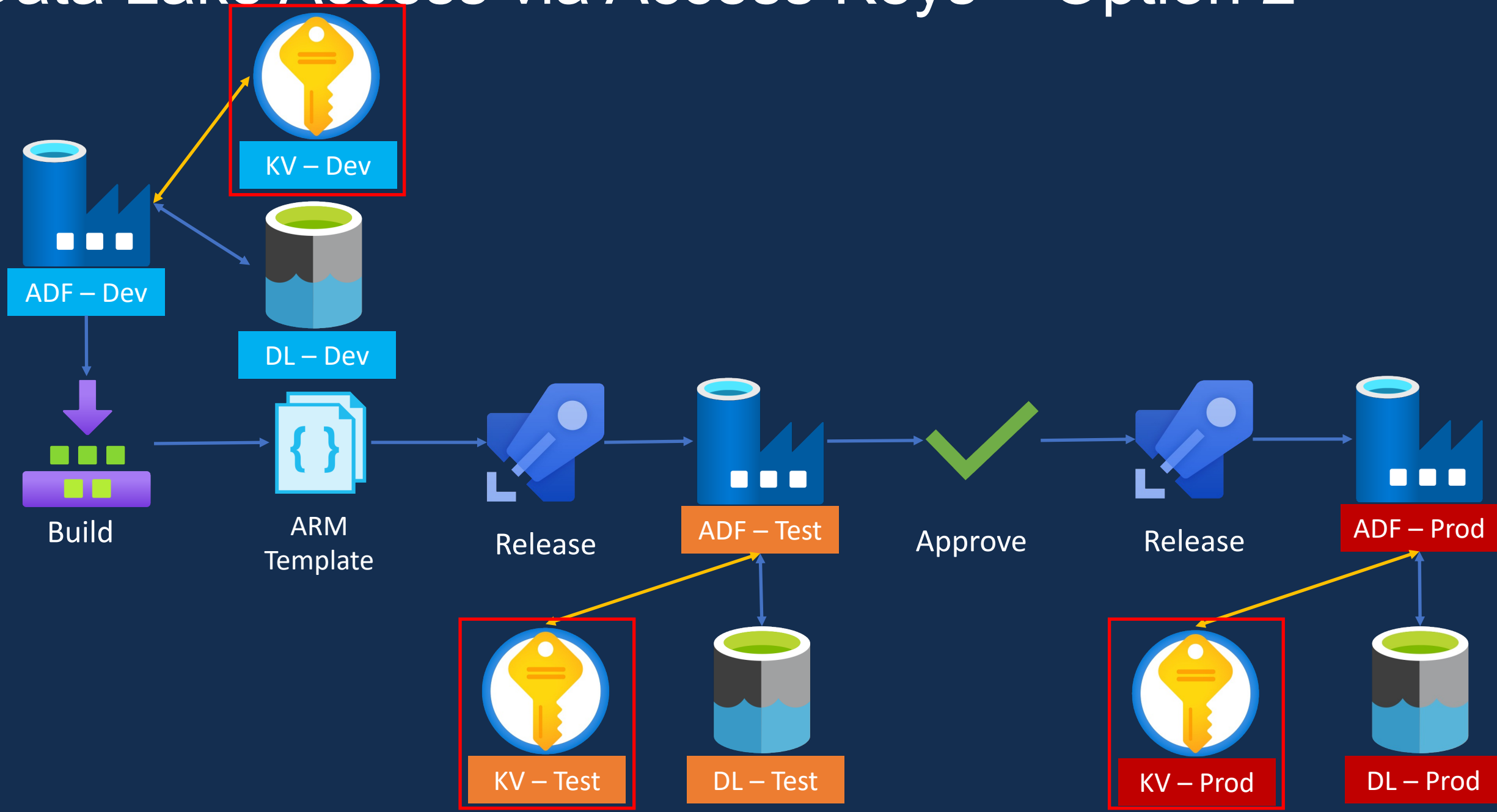
# Data Lake Access via Access Keys – Option 1



# Data Lake Access via Access Keys – Option 2



# Data Lake Access via Access Keys – Option 2



# Key Vault Set-up

Env	Data Factory Name	Resource Group Name	Data Lake Name	Key Vault Name	GIT Enabled
dev	dev-ci-cd-demo-adf	dev-ci-cd-demo-rg	devcicddemodl	dev-ci-cd-demo-kv	Y
test	test-ci-cd-demo-adf	test-ci-cd-demo-rg	testcicddemodl	test-ci-cd-demo-kv	N
prod	prod-ci-cd-demo-adf	prod-ci-cd-demo-rg	prodcicddemodl	prod-ci-cd-demo-kv	N

Congratulations!  
&  
Thank you



Feedback

# Ratings & Review

Thank you  
&  
Good Luck!

# Version History